AD NUMBER

AD481148

NEW LIMITATION CHANGE

TO

Approved for public release, distribution unlimited

FROM

Distribution authorized to U.S. Gov't. agencies and their contractors; Administrative/Operational Use; Sep 1965. Other requests shall be referred to Office of Aerospace Research, Washington, DC.

AUTHORITY

AFOSR ltr, 29 Nov 1968

AFOSR 65-1425
SEPTEMBER 1965

AUTOMATIC INDEXING

from Machine Readable Abstracts of Scientific Documents

Pranas Zunde
DOCUMENTATION INCORPORATED
BETHESDA, MARYLAND

OFFICE OF AEROSPACE RESEARCH
UNITED STATES AIR FORCE
WASHINGTON, D. C.

AFOSR 65-1425

SEPTEMBER 1965

AUTOMATIC INDEXING FROM MACHINE READABLE ABSTRACTS OF SCIENTIFIC DOCUMENTS

Pranas Zunde

DOCUMENTATION, INCORPORATED
BETHESDA, MD.

OFFICE OF AEROSPACE RESEARCH

UNITED STATES AIR FORCE

WASHINGTON, D.C.

In memoriam

DR. MORTIMER TAUBE

ACKNOWLEDGMENTS

The idea of automatic indexing of scientific abstracts by a method similar to FAST emerged early in 1963 in discussions with Dr. Vladimir Slamecka, then head of the Special Studies Division of Documentation, Inc.*) However, regular studies and systems design were started only in 1965 under the contract with the AFOSR. I greatfully acknowledge valuable comments and suggestions of Dr. Mortimer Taube, Alexander Kreithen, Jack J. Wolfire, Wolf Kuebler, Philip Dressler and other staff members at Documentation, Inc., which they readily contributed during various phases of the development of the project. The FAST programs for IBM 1410 computer were written by Dillon Scofield. It is also my great pleasure to acknowledge the outstanding editing work of Richard G. Katz and John A. Linford, both Senior Information Specialists at Documentation, Inc.

---

*) Now Dean of the School of Information Science, Georgia Institute of Technology, Atlanta, Georgia.

TABLE OF CONTENTS

INTRODUCTION

The work described in this report, which includes both basic research on automatic indexing and the design of an operational system, was performed at Documentation Incorporated under the contract with the Air Force Office of Aerospace Research, No. AF 49(604)-4236. Phase one of the project implementation was the preparation of a state-of-the-art survey and a bibliography, which are published as Part I of the report. It includes a thorough evaluation of all the reported experiences and results in the automatic indexing field. Phase two of the project was a detailed analysis of the particular characteristics of the input material for which the automatic indexing system was to be designed. Mathematical models for certain index formation processes were derived. The results of the findings are described in Part II of the report, which also contains the description of the proposed Formal Auto-indexing of Scientific Texts (FAST) System. On June 30, 1965, the Air Force Office of Scientific Research invited a selected audience of representatives of government agencies and non-government organizations with vested interests in information processing field to a demonstration of this new FAST system at the Documentation Incorporated premises in Bethesda, Md. The opening remarks of Col. Donald R. Currier of the AFOSR and of Dr. Mortimer Taube of the Documentation Incorporated follow this introduction.

STATEMENT OF COL. DONALD R. CURRIER

This subject indexing work which has been done under the ILSE contract is an example of what happens when the time has arrived for a good idea to come to reality. It is not because of some miraculous technical breakthrough that we have a demonstrable system today although Mr. Zunde here at Documentation Incorporated has pushed the state-of-the-art forward a significant notch. It is because the one missing ingredient in most previous experiments with computer indexing was present this time. This ingredient was a very large store of abstracts that not only had to be put in machine useable form, but also had to be hand subject indexed to meet a basic ILSE requirement for a controlled vocabulary for subject searches. All of the costs to do the above tasks could be considered as "sunk costs" from the standpoint of the automatic indexing task. They would be incurred anyway even if no automatic indexing research were to be done. Thus, a tested working media for the next step was all paid for.

Some extra money came about because the original estimates someone made of what the DOD portion of the ILSE effort would cost were high. The money fell into my hands just at the time when I had become interested in adding this sort of capability to MCDS and had been discussing the matter with the people at Documentation Incorporated. It was not difficult to sell the idea to D. Frese, the ILSE Panel Chairman, that we might save a lot of future ILSE money by risking some of the current years surplus nor to convince him that a modest extension to

test the general applicability using OAR data in other areas of science was worthwhile from the DOD standpoint.

This work has the potential to save the government a great deal of money and people's time if it is applied. More importantly, it may be the key to the precisely directed exchange of one type of scientific information on a scale that has not been possible before anywhere.

I would now like Dr. Taube, Chairman of the Board of Documentation Incorporated and a man with considerable experience in information retrieval to set the stage for the presentation by Mr. Zunde.

STATEMENT OF DR. MORTIMER TAUBE

As Colonel Currier has pointed out, we were able to begin this automatic indexing project without the necessity of investing in the input costs for a data base. This permitted us to concentrate on the logic of the indexing problem. Following our usual procedure, in order to avoid re-inventing the wheel, we did a complete study of the existing literature on automatic indexing. Out of this study there emerged the conviction that many organizations who have preceded us in this area, have restricted themselves to speculating on the number of different ways to do the job, rather than on the basic question of determining whether or not automatic indexing was indeed feasible and could be accomplished with existing equipment and program capability.

We discovered in this area, as in many others, a tendency on the part of those who speculate and are not concerned with the solution of operating problems to complicate the problems more than is necessary. One can devise many methods for selecting a proper set of index terms from a machine-readable text. The problem is to determine the simplest and most economical method which will create a usable index of high quality. In this field, as in many others, it is a conviction of Documentation Incorporated that the simplest system which works is the best system for any particular application.

Documentation Incorporated is internationally known for the development of coordinate indexing which is now standard operating procedure with all organizations using manual indexing with computer manipulation of the index. Coordinate Indexing is based on the bet that indexing can be accomplished with a set of terms with relations among the terms limited to Boolean functions of "and,""or," and "not." Many people have proposed adding much more complex relational systems, but in no case has it been proved that such complexity does more than raise the cost without improving the system. We are aware that in a Boolean system, we may not be able to distinguish between venetian blinds and blind Venetians. But we will only worry about this problem if we are certain that in our system of information we have stored an equal amount of data on both blind Venetians and venetian blinds. If we have only information on building materials, namely venetian blinds, we will not worry about the possibility of retrieving information on blind Venetians if there is no information on blind Venetians in the system.

Now it turns out to be the case that many people who have developed elaborate syntactical and semantic rules for automatic indexing, have done so without regard to the actual amount of noise or erroneous information which might be retrieved with simpler and less costly systems; therefore we have followed information theory and have tried to create the freest and simplest system consistent with the creation of an index

adequate for the uses to which it will be put. Mr. Zunde will tell you about the details of this system. We are not claiming a breakthrough or any great discovery in this regard, but merely another demonstration that rigorous, logical analysis and attention to the requirements of theory and economic feasibility can deliver important and usable operating answers.

PART I

STATE-OF-THE-ART OF MACHINE INDEXING

I.1. SCOPE OF THE STATE-OF-THE-ART STUDY

The advent of computers has opened new vistas in the information processing field. Among the many areas which have already received some consideration has been the mechanization of indexing. Most likely it will result in the elimination of much human effort from the indexing process, with the reduction of human bias or distortion from the process as a secondary effect.

This state-of-the-art study briefly surveys recent developments in the machine or automatic indexing field. At the present time, automatic indexing is basically in an experimental stage. Various methods of automatic indexing are described and evaluated. Areas of research required to improve operational qualities of proposed systems are indicated. It is hoped that this study will help systematize the thoughts of persons interested in automatic indexing and that it will suggest various possible approaches to solutions of their particular problem.

Emphasis has been placed on quantitative rather than qualitative methods of automatic indexing. At this stage of development quantitative methods offer much greater possibilities for practical application because they are less complicated and therefore less expensive and time consuming. Qualitative methods, such as the methods of linguistic analysis which form the basis of machine translation, were only remotedly considered by a few researchers for application in machine indexing and very little has been done to test these ideas in practical experiments.

The study also does not cover research aimed at full text searches of documents, even though there are some problems common to index generation. It was not considered the purpose of this study to investigate that which makes indexing necessary or superfluous, but how to produce an index by machine.

The assumption is made throughout that material to be processed is in machine readable form. In other words, the study neither concerns itself with the conversion to machine readable form nor with the equipment required to perform the conversion. It is realized that at this time conversion to machine readable form solely for the purpose of machine indexing would not be economical. In the near future, however, print-reading devices such as computers with optical scanners should be sufficiently developed to make this task economically feasible and desirable. For material not yet printed, type-punching devices attached to typewriters and type setting machines could readily produce machine readable records as by-products. It is therefore anticipated that the time is not too far off when recording information directly in machine readable form will be a common thing. This could then open the doors for a wide-scale application of machine indexing - and machine abstracting - systems.

Note: As this state-of-the-art study was being completed, the author received a published copy of a similar study by Marie E. Stevens (54). Fortunately, there seems to be no real duplication of effort. Whereas the work of M. E. Stevens covers a wider area of the utilization of machines in indexing, this paper is more task oriented towards system design.

I.2. MACHINE INDEXING METHODS

For the purpose of investigating automatic indexing, it is convenient to differentiate between indexing by extraction and indexing by assignment. In the first case, viz. indexing by extraction, selected words which appear in the documents are used as indexing terms. The design objective is to make the machine select words which adequately represent the contents of the document and to record them. In the case of indexing by assignment, decision is first made by the programmed machine as to which particular category or class of human knowledge the document to be indexed belongs and then words, which are considered to be most pertinent descriptors of that particular class or category, are assigned as indexing terms. These words may or may not appear in the document itself. Thus, if the document is on: INVESTIGATION OF TURBULENCE EFFECTS IN IONIZED PLASMA FLOW, the derived indexing terms might be TURBULENCE, IONIZED, PLASMA, FLOW. The assigned indexing terms might be, for instance, MAGNETOHYDRODYNAMICS and PHYSICS. Obviously, the second method can also be referred to as automatic categorization.

I.2.1. Indexing by Extraction

One of the crucial problems in selecting and extracting indexing terms from the text of the document is to find the significant ones, viz. such terms which would most adequately represent the contents of the document for their later identification in a retrieval process. There are

several criteria which can be applied, and which have been more or less successfully applied, in selecting significant words from the text. These criteria may be classified into four main categories:

positional and typographical criteria

semantic and syntactic criteria

pragmatic criteria

statistical criteria

Positional and Typographical Criteria. Significance is often attributed to words in titles of the documents or in section headings. On a sample of 25 articles, included both in Physics Abstracts and Chemical Abstracts, Maizell (183) showed that the titles alone contained about 50-70 percent of the key terms under which the articles were actually indexed. A study by Montgomery and Swanson (195) of the Index Medicus led them to the conclusion that titles alone provide about 50 percent of clues for judging the relevance of a given article to a given information need.

A well known operational system based on this concept is the KWIC (Key Words In Context) index of significant words in titles, which is being used for Biological Abstracts and a number of other indexes.

Baxendale (9) proposed to partition the title into phrases of three types: prepositional phrases, phrases containing a conjunction, and clauses. The identification of specific structural features within a

title is aided by a dictionary of approximately 300 entries consisting of the letters of the alphabet, certain punctuation symbols, and certain words representing relatively stable syntactic features such as auxiliary verbs or irregular adverbs. The eligible index terms, one-, two- or three-words long, are recognized and selected by the computer from the partitioned title units, and their grammatical function, such as adjective or noun, is then assigned. Thus, the selection and assignment rules are based on the position of the words rather than by the recognition of their grammatical function. The computer program for this system, written in the COMIT language, is called "Title Analyzer."

There are other positional criteria besides titles. According to Baxendale (7), references on composition techniques state that the "strategic" location for the prime thought of a paragraph is either first or last. In other words, these are the positions for the greatest emphasis. An investigation of a sample of 200 paragraphs corroborated the rule: in 85 percent of the paragraphs the topic sentence was the initial sentence and in 7 percent the final. Operating on these sentences only not only would greatly reduce the volume of the article, but also would have the added advantage of eliminating much of the less significant vocabulary as well as many of the least pertinent parts of speech, such as verbs and adverbs. Baxendale reported in her experiments the percentage of condensation achieved by selection of topic sentences and deletion of common words ranging from 6.3 to 18.9 percent or an average of 11.6 percent.

Two quasi-automatic methods of indexing proper nouns (quasi-automatic because they involve a considerable amount of human postediting) were described by Artandi (4). Both methods are based on the criteria that proper names appear capitalized in natural text.

Semantic and Syntactic Criteria. Significance is contributed to words in virtue of their relation to certain other words, also called cue words, such as "summary,""conclusion," etc. A technique utilizing this method is described in the Ramo-Wooldridge report on automatic indexing and abstracting (48). A cue word glossary is compiled for the population of documents. By hypothesis the cue words tend to indicate (or appear in proximity to) important or significant material. Using the cue words, an initial set of sentences is selected, which is then examined by the program to identify those words which are the most likely to be the key words of the document. The cue word list also contains common words, which carry little or no information, but these common words are assigned a weight of zero and are thereby eliminated from the document. Thus, every word of text is classified as either cue word, insignificant word, or potential key word. The immediate application of key words is using them as indexing terms.

O'Connor (45) studied the cue- and key-word method by searching for computer rules which would duplicate indexing done by subject specialists for a pharmaceuticals retrieval system. To begin with, he investigated just

a single term toxicity. One hundred documents, containing thirteen toxicity papers, was the first random sample from the total population of some ten thousand documents in the Merck Sharp and Dohme Research Center Library. Computer-generated word frequency lists were prepared for each sample document. A thesaurus group of likely toxicity keywords was derived from the retrieval system's indexing guide, a medical dictionary, and the papers in the sample. Thirty sample papers each contained at least one keyword; eight of these were papers on toxicity. Five other papers on toxicity appeared to contain no keywords. Frequencies and positions of keywords in documents, and differing weights for keywords, were used in the attempt to reduce keyword overassigning of toxicity. Frequencies did not appear to help. To some extent, weighting helped but the best criterion seemed to be occurrence in summaries.

A further investigation of this approach lead to many other expressions for toxicity cues, but they could not be used directly for mechanized indexing because they were unlikely to recur in other papers on toxicity. Study of these expressions suggested their generalization to "expression forms" containing variables. The possible values of the variables were defined for computer use by lists of "substance-contact words" and "disorder words." "Expression forms" permitted assigning the indexing term toxicity mechanically to four relevant papers which contained no original keywords. Various elaborate indexing rules using "expression forms" were suggested. The best of these, combined with rules involving

keywords, selected all twenty-one papers on toxicity as well as nine irrelevant papers from the given sample.

We might also include in this category Baxendale's (7) suggestion to select prepositional phrases as containing the significant words of the text. According to Baxendale, a phrase is likely to reflect the content of an article more closely than any other simple construction. Therefore, she proposed to make the preposition itself the indicator for initiating selection of index units. The length of a phrase varies from two to seven words, with an average of four words (based on a count of words per phrase in 350 phrases). Thus, "by running the risk of selecting too large or too small a unit, but obviating the necessity of discriminating to select nouns and their modifiers, it is possible to program a computer to recognize the preposition by table look-up and then automatically select the next four words unless a second preposition or a punctuation mark is encountered." For example, the machine would select the underlined words or work groups in the following sentence: Within the scope of natural English language, an infinite number of different sentence structures is possible. The percentage of condensation achieved by selection of prepositional phrase and deletion of common words was reported from 4.8 to 18.2, the average being 11.3 percent.

Pragmatic Criteria. This approach is based on the assumption, as proffered by Artandi (4, 5, 6) and Kraft (28), that it is possible to

create a vocabulary or a list of terms and a syndetic apparatus for a given subject area which, if sufficiently representative of the field, may be used in the construction of indexes to materials in the same subject area by matching the thesaurus against the text of documents. A system based on this criteria alone was described by Artandi (6). The vocabulary of the proposed system includes the following elements: (1) terms in the detection part of the vocabulary, each of which may consist of one or several words, entirely identical with the phraseology of the text; (2) terms in the expression part of the vocabulary, which are the terms of the final index and may or may not be identical with the corresponding detection term. A section of a chemistry textbook was selected as the experimental document and it was reported that the vocabulary of the system contained 744 detection terms. Unfortunately, the report does not contain information on the size of the document or the total number of words it contained, neither does it give a description of how the detection terms were derived or selected.

Kraft (28) describes in his paper a system claimed to be the first Selective Dissemination of Information (SDI) system in operation. The system includes an automatic indexing phase based on a similar approach to the one described above. The punched cards containing the abstracts are automatically indexed by the SDI program on the IBM 1401. The indexing can be done in either of two ways: (1) terms may be selected from the

abstract, title, and author's name if they do not match a word on an exclusion list of common words stored on magnetic tape; (2) terms may be selected from the abstract, title, and author's name if they match a word in a dictionary stored on magnetic tape and are not on an exclusion list of common words. The manually-selected descriptors are also indexed by the program. Using the dictionary approach combined with the exclusion list, an average of 22 keywords are chosen per item. The exclusion list technique alone indexes an item by an average of 41 keywords. It must be noted, however, that the requirements for that system, which serves salesmen and system engineers of the IBM Corporation Midwestern Region Office in Chicago, Illinois, are not very sophisticated.

Statistical Criteria. Statistical approach to automatic indexing seems to be the most promising. Luhn, Baxendale, Levery, Williams, and others have experimented with this approach. In most cases, the first step is deletion of insignificant words. This is done by designing a look-up list for the computer which might include pronouns, articles, conjunctions, conjunctive adverbs, copula and auxiliary verbs, quantitative adjectives and similar words. The size of such a list varies from 100 to 700 words for the systems reported. Condensation thus achieved ranges from 50 to 70 percent (7). A modified procedure is to delete all words with three or fewer characters (6).

At this point, one approach, originated by Luhn (32), consists of making absolute frequency count of the remaining words, ordering them by

descending frequency, and selecting the words within a certain frequency range as the most significant ones (7). The justification of measuring word significance by use-frequency is based on the fact that a writer normally repeats certain words as he advances or varies his arguments and as he elaborates on an aspect of a subject. No effort is made to differentiate between word forms. Luhn argued that within a technical discussion there is a very small probability that a given word is used to reflect more than one notion. The probability is also small that an author will use different words to reflect the same notion. Even if the author makes a reasonable effort to select synonyms for stylistic reasons, he soon runs out of legitimate alternatives and falls into repetition if the notion being expressed was potentially significant in the first place.

As to the upper bound of the frequency range, Luhn proposed two solutions. One solution would be not to set any upper limit, and to eliminate the common words, which can naturally be expected to cluster in the high frequency region, by comparing them with a stored common word list. Another solution is to determine a high frequency cutoff through statistical methods to establish "confidence limits." Since degree of frequency has been proposed as a criterion, a lower boundary would also be established to bracket the portion of the spectrum that would contain the most useful range of words. The optimum locations for these cutoffs would be established from experiments with large samples of input data.

Luhn believed that it should even be possible to adjust these locations to alter the characteristics of the output. If non-common words fall into the high-frequency region, it would indicate their loss of discriminatory power. Common words falling in the region of acceptable frequency would be tolerated because of their lesser degree of interference. Thus, it may be anticipated that the cutoff line, once established, may be stable over many different degrees of specialization within a field, or even over many different fields.

In the experiments reported by Luhn, the determination of this frequency range was arbitrary. Luhn assumed that 10 to 24 of the highest ranking words are the most significant ones for document identification, 16 such words being the likely average.* The size of the document collection for which this size pattern would suffice has not been determined. Indications are, however, that size of collection is not a major function in determining optimum pattern size.

The refinements of this method are the "normalization" of the list viz. combining the terms on the list to notions by look-up in the special thesaurus, and switching to the so called "multi-dimensional patterns." For the latter purpose, the automatic process would proceed

---

* Baxendale (7) assumed the number of allowable words for the index as 0.5 percent of the total number of words in the article, the ones which occurred with the highest frequency after the deletion of common words.

to extract from the sentences all word pairs consisting either of two adjoining first order words or of a first order word coupled to a second order word, the first order words marked by an appropriate sign. A record is then developed, giving for each first order word (node) all the words which have been found paired to it (branches).

Instead of operating with single words, Meetham (191) investigated the possibility of extracting significant word pairs and word groups for an automatic generation of descriptor systems and for indexing. All possible pairs of words were examined and those pairs selected which occur so frequently in the same document (in relation to their frequencies of occurring separately) that the frequency of their co-occurrence is probably not by chance. The second step is to discover word-groups from an examination of the word-pairs. The words from which such groups are made are picked out from a word list by using a word-word binary matrix to represent the association between pairs of words.

A relative frequency approach proposed by Edmundson and Wyllys (23) takes into account the fact that, according to information theory, a word's information value should vary inversely rather than directly with its frequency, its low probability evidencing greater selectivity, or deliberation, in its use. It is the rare, special, or technical word that will indicate most strongly the subject of the author's discussion. Here,

however, by "rare" is meant rare in general usage, not rare within the document itself. In other words, Edmundson and Wyllys claim that it is wrong to treat a document as the universe of words. Rather, the frequency of a word in a document should be compared with the frequency of the same word in general use, viz. to regard the contrast between the word's relative frequency f within the document and its relative frequency r in general use as a more revealing indication of the word's value in indicating the subject matter of document d. Four types of significance functions s(f,r) are proposed, $s = f - r$, $s = f/r$, $s = f/(f+r)$, and $s = \log(f/r)$, of which $s = f - r$ or $s = f/r$ are suggested as the best choice. According to the authors, defining significance in terms of the contrast between frequency in a document and in general usage would give low significance both to normally rare words which occur rarely in the document and the common words used frequently within the document itself. The relative frequencies are calculated as follows:

$$f_{wd} = N_{wd}/N_d \qquad r_{wc} = N_{wc}/N_c$$

where

N_{wd} is the number of occurrences of word w in document d

N_d is the total number of running words in d, i.e. $N_d = \sum_w N_{wd}$

N_{wc} is the number of occurrences of word w in the class of documents c

$N_c = \sum_w N_{wc}$

A further refinement of the process of automatic analysis would be the development of special sets of reference frequencies for special fields of interest. Two benefits are claimed for this: it would become

possible to classify documents as to field, and it would become possible to note the significance of words which are frequent in a very large reference class c_o of literature (i.e. these words would be significant with respect to c_o) but which are rare in the special field.

To demonstrate how this method would operate, assume that the relative frequencies of m words have been established, both for a large reference class c_o of literature and also for n special fields of interest c_j, $j = 1,2,.....,n$. Thus, there would be $n + 1$ values of relative frequency for each word w, where w runs from 1 to m, and where

r_{wo} = relative frequency of word w with respect to the class c_o literature

r_{wj} = relative frequency of word w with respect to special field c_j.

Next, the $m \times (n + 1)$ matrix (r_{wj}) is formed, each column of which contains the frequencies of all the listed words for a particular field (the whole body of literature being represented in the first column) and each row of which contains the frequencies of a particular word in all the listed fields.

The Automatic indexing would then proceed as follows: first, the determination of the words that are significant with respect to general literature by the comparison of the relative frequencies f_{wd} of words in d with the relative frequencies in the first column of the matrix (r_{wj});

second, the comparison of the document's frequencies with the other columns in the matrix in order to determine which column forms the "best fit" with the document; and third, the determination of the words that are significant with respect to the special field. One standard method for determining the "best fit" would be to find the column j whose frequencies differ least from those of the document. Once frequency-ordered indexes have been established for various subject-fields the automatic index of any new document can be compared with them by machine processes. According to the authors, the results of the comparison would determine, first, the subject field to which the document properly belongs (classification); second, other subject-fields with which it should be associated (cross-reference); and finally, those terms which are significant enough to be used as identification tags for the process of recovering the document (retrieval).

As an extension of the relative frequency approach, involving syntactic and semantic approaches, the author proposes the introduction of weighted frequency. The machine can be instructed to recognize the title by position and capitalization and to place a "title indication" after each word appearing in the title as it compiled its list. Similarly, it can place "first-paragraph indications" after all words it meets until it recognizes the end of the first paragraph. Every heading or sub-title can be tested for the words "summary" or "conclusions" and place a "summary indication" after each word in the summary paragraphs. At the conclusion of its "reading" of the article, the machine can compute

each word's weighted significance S according to the formula:

$$S = b_1 b_2 b_3 \, s(f,r),$$

where for a given word w,

$$b_1 = \begin{cases} b_t & \text{if w bears a title indication} \\ 1 & \text{otherwise} \end{cases}$$

$$b_2 = \begin{cases} b_p & \text{if w bears a first-paragraph indication} \\ 1 & \text{otherwise} \end{cases}$$

$$b_3 = \begin{cases} b_s & \text{if w bears a summary indication} \\ 1 & \text{otherwise} \end{cases}$$

and where b_t, b_p, and b_s are preassigned weights, all greater than one, for occurrence in title, first paragraph, and summary, respectively.

Alternatively, statistical methods of this type might be used as preliminary sorting for later application of non-statistical criteria. For example, when a word already known to be somewhat significant by statistical methods also occurs in the title, its significance might be taken as guaranteed, and the machine program could recognize the fact by placing it on the "definitely significant" list, even though the word was outranked in significance by other words. Recapitulating, the final selection of significant words would be based on three criteria: (1) significance of the word with respect to general literature, (2) significance of the word with respect to a specialized field, and (3) placement of the word on a "definitely significant" list. Under criteria 1 and 2 there would be an

alternative of selecting either all words whose significance value exceeded a predetermined threshold value s, or only the first n words in order of significance from the highest down, adding, in either case, those words selected by criterion 3.

A somewhat similar but simplified technique is described by F. Levery (30) of the International Business Machine Corp. in France. Non-common words are first combined to form notions with the help of a dictionary of synonyms, and the frequency of the notions is counted for selection of significant terms. Two criteria were applied for the selection of keywords: (1) frequency of the appearance of notions above the average frequency of all notions in the text studied, and (2) the frequency of the appearance of the word to exceed the average frequency in the entire collection. The experiments were conducted on French language texts dealing with the manufacture and study of glass. Thirty documents were machine indexed, each document being 200 to 600 words long. The total number of words was 10,721. The deleted list for the whole collection consisted of 668 words, which appeared 6,589 times and thus accounted for 61.4 percent of the words present. The 1,681 different non-common words found in the collection were grouped into 897 notions. The 30 most frequent notions accounted for over one-fourth of the non-common words appearing (4,132). The input processing was done on an IBM 7094 computer which supplied for each document a word list in alphabetic order and another list in order of frequency.

The technique for selecting significant words, proposed by Oswald (47), has the following main features: (1) Insignificant viz. common words are deleted and only words that are significant in the context of the document are retained. (2) The retained words are frequency counted. (3) Next, every juxtaposition (of two or more words) involving a high-frequency word is recorded as a significant word group. The recording of such groups begins with those that contain the single word of highest frequency and continues until six successive Uniterm words, in order of descending frequency on the Uniterm frequency list, produce either no significant groups or no new significant groups. This rule produces auto-indexes whose lengths, although differing, usually lie within the limits of 1 to 3 percent of the total vocabulary of any given article.

Finally, special consideration should be given to the text condensation and index editing method by consolidating concept related words which are spelled in the same way at their beginning, such as elliptic and ellipticity. The procedure proposed by Luhn (32) is a statistical analysis routine consisting of a letter-by-letter comparison of pairs of succeeding words in the alphabetized list. From the point where letters failed to coincide a combined count was taken of the non-similar subsequent letters of both words. When this count was six or below, the words were assumed to be similar notions; above six, different notions. Although this method of word consolidation is not infallible, errors up to 5 percent did not seem to affect the final results.

1.2.2. Indexing by Assignment

This type of indexing presupposes categorization or classification of documents as the first step in the selection of indexing terms. Various approaches to automatic document categorization will be briefly surveyed here.

Maron's (36) method starts with selecting statistically cue words from a sample population of documents previously assigned to certain categories by human indexers. The complete corpus consisted of 405 different documents and was divided into two groups. Group 1 contained 260 abstracts which appeared in the March and June issues of the 1959 IRE Transactions on Electronic Computers, and was the basis for the statistical data necessary to make the subsequent predictions. Group 2 consisted of 145 abstracts which appeared in the September 1959 issue of the Transactions and was used to test the system.

A classification system of 32 categories was created similar to, but not identical with, the classification system used in the IRE Transactions, and each one of the 260 documents of Group 1 was carefully read and "sorted" into one or more of the categories. In the majority of instances a document was indexed under a single category, but in about 20 percent of the cases a document was indexed under two categories, and in only a few cases under three categories. The highest number of documents in a single category was 37, and the lowest was 2.

Next, every word in each of the documents of Group 1 was key-punched. There was a total of over 20,000 word occurrences with an average of 79 words per document, and a total of 3,263 different words. The 55 most frequently occurring logical type viz. common words (e.g. the, of, a, etc.) accounted for 8,402 of the total (20,515) occurrences. Thus, less than 2 percent of the words accounted for over 40 percent of the total occurrences. They were rejected as candidates for cue words.

The most frequently occurring non-common words were considered next. This list contained words such as "computer,""system,""data," "machine," etc. They also were rejected as possible cue words because it was felt that they had little discriminating power to be cues for the specification of subject content within the general field of computers. Of the total 3,263 different words, 2,120 or 65% occurred less than three times in the 260 documents. They were also rejected as possible cue words because they were too specific (provided they were indicative of the contents of the document at all). This left just over 1,000 different words with neither a very high nor very low relative frequency of occurrence. A listing was made showing the number of times each of these 1,000 words occurred in the documents belonging to category 1, category 2, etc. Each word on the list was checked to determine whether or not it "peaked" in any of the 23 categories. If a word did peak it was felt that the word would be a good cue. If the distribution was flat for a given word, then it was rejected. An attempt was made to find at least one word to peak in

each of the 32 categories. In this way, 90 different words were finally selected as cue words.

Then the problem was conceived as follows: Given that a document, say D_1, contains one or more cue words W_i, what is the probability that D_1 belongs to each of the categories C_1, C_2, C_3, and so on. Maron used the well known Bayes prediction equation to calculate these probabilities. For one cue word W_i, the equation is:

$$P(C_j|W_i) = \frac{P(C_j) \cdot P(W_i|C_j)}{P(W_i)}$$

$P(C_j)$ is the so-called a priori probability that a document will be indexed under the j-th category and $P(W_i|C_j)$ is the probability that if a document is indexed under the j-th category it will contain word W_i. For any W_i, the denominator $P(W_i)$ is a constant and hence the equation may be rewritten as follows:

$$P(C_j|W_i) = k \cdot P(C_j) \cdot P(W_i|C_j)$$

where k is a scaling factor. The value of $P(C_j)$ is estimated by counting the number of index entries that are made under the j-th category and dividing this by the total number of index entries. The values of $P(W_i|C_j)$ are estimated by counting the number of occurrences of the i-th word which belong to documents that were indexed under the j-th category and dividing through by the total number of cue word occurrences in all documents belonging to the j-th category.

In the general case where a document contains different cue words, W_k, W_m, ... W_s, the probability that the document belongs to the j-th category is computed as follows:

$$P(W_k, W_m, \ldots\ldots W_s, C_j) = k \cdot P(C_j) \cdot P(C_j, W_k) \cdot P(C_j, W_m) \ \ldots \ P(C_j, W_s)$$

The values of the left hand side of the above equation are called "attribute numbers." Thus, 32 attribute numbers are obtained for each document, one for each of the 32 categories.

It turned out that in the initial group of 260 documents, 12 documents contained none of the 90 cue words, and hence no automatic indexing was possible for these 12 documents. Also there was an error preventing one of the remaining documents from being automatically indexed. This left 247 documents. In 209 of the 247 cases (84.6%), the category with the greatest attribute number in each output list was a correct category. If the document had at least two cue words, then the probability that the category with the greatest attribute number is a correct one was 91 percent. In Group 2, which was the new input to be tested, of a total of 145 documents, 20 contained no cue words, and 40 contained only one cue word. This left 85 documents, each containing at least two different cue words. In 44 (51.8%) of these 85 cases the machine printed the correct category at the top of the output list, i.e. the category with the greatest attribute number was the correct category. The probability that the machine will print out the correct category in one of the first three positions was 80 percent.

A modified approach to evaluate the "goodness" of the cue words was proposed by Trachtenberg (59). It involves calculating for each potential predictor or cue word (a) the non-correlation factor of word occurrence category, or the uncertainty of category given the occurrence of a word W_i in a document

$$H_i = -\sum_j p_{ij} \log p_{ij} \qquad 0 \leq H_i \leq \log k$$

where p_{ij} is the probability that a document with the word W_i falls into the category C_j,

and (b) a special measure involving the log of the ratio of the a posteriori to the a priori probability, viz.

$$M_i = \sum_j p_{ij} \log \frac{p_{ij}}{p_j}$$

A word that has a high value for M_i and a low value for H_i would be selected as the cue word.

Similar procedures were proposed to treat word frequency information. The corresponding equations are:

$$H_i (f_s) = -\sum_j p_{ij} (f_s) \log p_{ij}(f_s)$$

$$M_i (f_s) = \sum p_{ij} (f_s) \log \frac{p_{ij} (f_s)}{p_j}$$

where f_s is the range of the values of relative frequency of a word appearing in a document to the total number of words in that document, and $p_{ij}(f_s)$ is the probability that the document falls in category C_j given that the

relative frequency of word W_i in the document is in the interval f_s. No testing of the proposed method was made.

Borko (11) proposed a method which uses "factor loadings" of terms as probability measures for determining the category to which a document belongs. Briefly, his approach is as follows. Six hundred and eighteen psychological abstracts were coded in machine language for computer processing. The total text consisted of approximately 50,000 words, of which nearly 6,800 were unique. The computer program arranged these words in order of frequency of occurrence. From the list of words which occurred 20 or more times, excluding syntactical terms such as, and, but, of, etc., the investigator selected 90 words for use as index terms. These were arranged in a data matrix with the terms on the horizontal and the document number on the vertical axis; the cells contained the number of times the term was used in the document. Based on these data, a correlation matrix, 90 by 90 in size, was computed which showed the relationship of each term to every other term. To compute the correlation coefficient from raw score data (Document-Term Matrix), the following formula was used:

$$r_{xy} = \frac{N\Sigma XY - (\Sigma X)\cdot(\Sigma Y)}{\sqrt{\left[N\Sigma X^2 - (\Sigma X)^2\right]\left[N\Sigma Y^2 - (\Sigma Y)^2\right]}}$$

where N = total number of documents, and X and Y are terms being correlated. A computer program for calculating these correlations was written by the Systems Development Corp.

There are a number of methods for estimating the commonality. The simplest procedure would be to choose the highest correlation coefficient from among the other correlations in that set. By grouping together the related terms, a classification system for the given corpus of documents could be derived. However, this is not a task that can be done by inspection. In the 90 by 90 matrix, which is symmetrical, there are 4,005 correlations. In order to analyze the data in a precise fashion, Borko employed the technique of factor analysis.

The purpose of factor analysis is to reduce the original correlation matrix to a smaller number of factors. A factor corresponds to the eigenvector. The size of the eigenvector, i.e., the eigenvalue, is equal to the contribution of the variance made by that factor. The first eigenvector, or factor, accounts for a relatively large proportion of information and each succeeding factor accounts for less. In factor analysis it is not necessary to account for the total variance of the correlation matrix, for it is known that a certain proportion of the variance is unique or specific to the given set of documents in the experimental situation. It is the common variance which is of interest only, viz. that portion of the variance that is due to the relationship among the terms and which would continue to be true for all sets of documents. The problem, of course, is to determine the proportion of the total variance which is common.

Based on the above considerations, the matrix was factor analyzed and the first ten eigenvectors were selected as factors. These were rotated for meaning and interpreted as major categories in a classification system. These factors were compared with, and shown to be compatible with but not identical to, the classification system used by the American Psychological Association.

A similar approach to the problem solution chosen by Maron is reported by Williams of the IBM Corporation (61). He also proposed a discriminant coefficient to identify significant words. This discriminant coefficient is a function of the relative frequencies of the i-th word in the j-th category

$$\lambda_i = \sum_j^n \frac{(p_{ij} - \bar{p}_{ij})^2}{\bar{p}_{ij}}$$

where

$p_{ij} = \dfrac{f_{ij}}{\sum_i^m f_{ij}}$ is the relative frequency of the i-th word in the j-th category

and $\bar{p}_{ij} = \dfrac{1}{n} \sum_j^n p_{ij}$ is the mean relative frequency per category of the i-th word.

These coefficients are calculated from the data obtained from a small set of reference documents previously classified into categories

(hierarchical classification structure assumed) by human indexers.[1] The discriminant coefficients thus computed are used to set up discriminant thresholds determining which words will be used in the classification equation and to assign weighting factors to the words themselves. The computer program categorizes documents by comparing the observed with the theoretical word frequencies and computing a Relevance Value (RV) for each document with respect to each category. The RV equation is

$$RV_j = 1 - \left[\frac{.01}{m} \sum_{i}^{m} \lambda_i \frac{(p_{io} - p^*_{ij})^2}{p^*_{ij}} \right]$$

where p_{io} is the relative observed frequency in the document, p^*_{ij} is the relative theoretical frequency of the i-th word in the j-th category after transformation to document size, and m is the number of word types in the group. Documents, which show highest RV for a particular category, are classified accordingly. Those documents having a RV outside the standard deviation limits would be returned for re-evaluation.

A somewhat simplified approach was taken by Stevens (55) of the National Bureau of Standards. The SADSACT (Self-Assigned Descriptors from

---

[1] For the experiment, 400 computer abstracts prepared and published by Cambridge Communications Corp. were selected. Each of the abstracts was classified by CCC in their normal operation. Three hundred of the 400 abstracts were used as reference documents, and were equally divided among the 20 categories of the classification system. The remaining 100 were used as the test documents. The objective of the experiment was to classify the 100 test documents into their correct categories.

Self and Cited Titles) method correlates descriptors or indexing terms with significant words in a representative sample for the population of documents to be indexed, viz. each significant word in the title and in the abstract of the document is associated with each of the descriptors previously assigned to that document. Descriptors that occurred three or more times in the 100-item sample were retained as "validated descriptors." For the validated descriptors, the word-descriptor association lists were then merged into a master vocabulary list which showed for each word the descriptors with which it co-occurred and the relative frequencies of its co-occurrence with each descriptor.

Thus, the SADSACT automatic indexing method used an ad hoc statistical association technique in which each word may be associated either appropriately or inappropriately with a number of different descriptors. The indexing procedure was carried out as follows. The text of the title of a new item and of titles cited as bibliographic references by the author was keypunched, and the byproduct punched paper tape was converted to cards for input to the computer. This input material was processed against the master vocabulary list to yield, for each word that matched a word in the vocabulary, a "descriptor-selection-score" value for each of the descriptors previously associated with that word. After all words from titles and cited titles were processed, the descriptor scores were summed and for some appropriate cutoff level, those descriptors having the highest scores were assigned to the new item.

The actual score value includes both a normalizing factor (based, for example, on the ratio of the number of previous co-occurrences of this word with a particular descriptor to the number of different words co-occurring with that descriptor) and a weighting formula that gives greater emphasis to words occurring in "self"-title (the authors own choice of terminology) than to those occurring in cited titles. Similarly, greater emphasis is given to words that coincide with the names of descriptors.

Baker (67) recognized the similarity between document classification and the problems inherent in the analysis of sociological questionnaire data and proposed the classification method based upon Lazarsfeld's (Stouffer) latent class analysis. Briefly, the latent class model assumes that the population - that is the number of documents in the sample - can be divided into a number of mutually exclusive classes. Usually the number of classes is determined by the investigator, although it is conceivable that this parameter can be determined mathematically. One starts by selecting the key words which characterize each class of documents. Then latent class analysis is used to compute the probability that a document having a certain pattern of key words belongs to a given class. For instance, assume that there are 1,000 documents in a file. These documents are to be classified into two classes - those dealing with computer automated instruction and those not directly related to this topic. The following key words are

selected in the search request:

1. computer
2. automated
3. teaching
4. devices

Each of the 1,000 documents is then analyzed to determine whether it contains one or more of the four terms. Sixteen (2^4) response patterns are possible, ranging from ++++ to 0000. A chi-square test enables one to estimate the latent structure from the observed data. Having obtained a latent structure which fits, one can compute an ordering ratio, which is the probability that a document having a given word pattern belongs to a particular latent class. For example, a document with all four key words present has a probability of .998 of belonging to class 1, i.e., it is concerned with computer automated instruction. The method seems to have merit, but no experiments were actually made to test it.

Obviously, once the document is delegated to a specific class or category by one of the above described methods, indexing terms or terms identifying the contents of that class can be tagged or assigned to the document.

I.3. MACHINE INDEXING EVALUATION

No absolute standards have been as yet discovered for machine indexing evaluation and measuring its "goodness" just as there are no standards and absolute measures of "goodness" of human indexing. Therefore some authors represent the viewpoint that until such standards and measures are discovered, if they can be discovered at all, only relative or indirect evaluation is possible by comparing a particular method of machine indexing with other operational systems, human or machine. Thus, there are two possibilities: (1) comparing machine indexing with human indexing and (2) comparing one machine indexing method with another.

Most investigators have attempted to compare machine indexing with human indexing and less has been done in comparing machine indexing vs machine indexing. The reason for this might be that so far there are only a few experimental automatic indexing systems being operationally tested and there is very little data on their actual performance.

Another suggested approach to the evaluation problem is to determine the quality of indexing by evaluating the quality of retrieval. Meetham (191) indicates in his report on the proposed automatic indexing system that the evaluation of the system on 53 inquiries in a sample collection of documents produced an overall relevance ratio up to 0.33

and an overall recall ratio up to 0.38. Unfortunately, very little was reported on the test methodology and systems parameters. Such data would have greatly enhanced the value of this pioneering effort.

Another small scale experiment, containing elements of this approach, is described by Swanson (56). A collection of 100 articles was chosen as an experimental library, and each article in the collection was studied in the light of its possible relevance to each of 50 questions asked. All the articles were on nuclear physics. Furthermore, in order to compare the effectiveness of text searching by computer with more or less conventional methods, the experimental collection of articles was catalogued by means of a subject heading index designed for this particular field of science. Three methods of retrieval were employed: (1) "Conventional retrieval" based on the subject heading index with no machine procedures involved; (2) Retrieval based on specifications of words and phrases in disjunctive and conjunctive combinations without any other retrieval aids; (3) Search requests formulated as described in the second case but with the thesaurus-like word and phrase group list and the index thereto as retrieval aids. The results in terms of "percent of relevant material retrieved averaged over all requesters and all questions" were reported as follows: Test One - 38 percent; Test Two -68 percent; Test Three - 86 percent.

No other practical studies in evaluating automatic indexing by retrieval efficiency besides these two limited size experiments are known.

Theoretically the possibility of such an evaluation and the implications of this method are discussed by O'Connor in his paper Mechanized Indexing Methods and Their Testing (44), which covers also a wide range of other problems related to machine indexing.

The problem of relevance, which involves high subjective criteria and therefore is hardly accessible to formalization, can be avoided by taking a strictly formal approach to index evaluation. In the case of human indexing, this would presuppose that the choice of indexing terms by one indexer is as good as by any other, provided of course, that the indexers are qualified specialists. Thus the choice of the indexing terms is accepted by the user at their "face value" within certain confidence limits, which are set by the variance of indexers in the selection of terms to tag a particular book or document. It is then up to the user, reference librarian or the information systems specialist to make the best use of the tools the indexer gives him to obtain maximum efficiency from the system subject to known limitations. The evaluation problem thus becomes a problem of a formal evaluation of the system as a communication channel, which on this basis is entirely accessible to mathematical analysis.

Extending this approach to automatic indexing, and in particular to indexing by extraction, we assume that the author of a book or document is competent enough to express the subject matter in pertinent words and that his choice of words is therefore accepted without questioning. All

the machine does is to eliminate from the author's text words which carry no information for any user, and to condense the meaningful terms to keep the index size within tolerable limits. Here again, the efficiency coefficient of information transmission would be the only formal measure in evaluating a given system.

I.3.1. Comparing Machine and Human Indexing

Our attempt to evaluate machine indexing by comparing it with a "well-reputed" human indexing system was part of a wider study on machine indexing and abstracting efficiency by Kurmey (29). The source data was obtained by selecting 50 abstracts at random from Chemical Abstracts, 10 each for the years 1951 through 1955, and tracing the abstracts to the original articles. These articles were then machine indexed, the only criterion of significance being the frequency of word appearance after the deletion of common words. Total number of words in all these articles was 131,283, number of different words (excluding common words) - 21,200, number of common words - 3,362, and average word frequency - 5.34. The predetermined frequency cutoff was obtained by dividing the list at the closest word frequency group corresponding to the number of terms assigned by Chemical Abstracts. The machine index so created was then used for direct comparison with the index entries assigned to the same article by Chemical Abstracts. In addition, the word frequency list for each article was manually scanned to determine if the terms assigned by Chemical Abstracts were at all present in the machine derived list.

Analysis of the index terms assigned by Chemical Abstracts was carried out manually prior to comparison with the machine index terms. For each article, the Chemical Abstract entries consisting of two or more words were broken into single word entries.

Comparison of the index words was carried out manually with two different approaches. In the first approach, the entire alphabetized non-common word list of an article was scanned to see if the word used by Chemical Abstracts was in the text of the article. The agreement between the Chemical Abstracts words and the word list was taken on a straight percentage basis. In the second approach, the number of words used by Chemical Abstracts was used as a cutoff to obtain words with the highest frequencies. The agreement was also taken on a straight percentage basis. The percentages use, as a base, the number of words in the Chemical Abstracts entries.

The average overall conformity between the alphabetized noncommon word list and Chemical Abstracts entries was found to be 81.76 percent. The average overall conformity between the subset of words of highest frequency and Chemical Abstracts entries was 27.63 percent.

It is apparent that "maximum-depth" indexing would cover most (81.76%) of the entries used in the Chemical Abstracts indexes for the articles. Most of the indexing terms used by Chemical Abstracts appear in the article hence the high agreement for the alphabetized noncommon words. However, the most frequently occurring words on the word frequency list

would only poorly duplicate human index entries for an article as only 27.63% concur with the entries in Chemical Abstracts. Therefore, Kurmey came to the conclusion that the subset of highest frequency words used in the article do not form adequate index entries for the article. Apart from constructing "maximum-depth" indexes consisting of all different words occurring in the article except those on an ad hoc "stoplist," Kurmey could not see any straightforward statistical method of arriving at index entries derived from a word frequency model of text with comparable entries in Chemical Abstracts. Possible improvement in the indexing entries was suggested by utilizing a thesaurus applicable to the field of chemistry to select significant words by direct match.

Contrary to Kurmey's results, comparing the lists of index terms obtained in the machine indexing experiments by relative frequency with the list of terms prepared manually, Levery (30) determined that on the average more than 85 percent of the keywords chosen by the analysts were also selected by the machine method. The lists prepared manually were arranged in descending order of significance of the keywords and the same words were obtained by automatic means. The elimination of common words and the regrouping of synonyms was done by hand.

In the related field of book indexes, Artandi (6), using a section of an inorganic chemistry textbook as the experimental document, compared the mechanical index with the average manually produced index found in inorganic chemistry textbooks. The author claims that the

mechanically produced index compares favorable in intellectual content with the average published manual index for the same type of material. Completeness of indexing (takes into consideration index entries actually assigned, incorrect entries, and omissions, that is, entries that should have been assigned but were missed for some reason), which is a numerical figure supposed to include both qualitative and quantitative evaluative criteria, was found to be practically identical for the experimental index and for the average of the published manual indexes checked. Entry density (the ratio of the total number of page references to the total number of pages) was 63.8% higher for the mechanical index than the corresponding average. Heading density (the ratio of the total number of index entries to the total number of words in the book) was found to be 8.8% lower than the corresponding average. The heading densities of the individual published indexes checked for the study fell in the range of 41.8% below and 56.0% above the average heading density value. It seems, however, that there might be a possibility of these figures being greatly biased because of predetermined matching instruction in the indexing procedures and rather artificial test conditions.

In a modified experiment by Artandi (4), two methods of machine indexing proper nouns were tested on the same inorganic chemistry textbook, which contained a total of 148 proper noun terms. Of the total of 324 entries produced by the machine, 208 entries or 63.1 of all the produced

entries were useless or not proper noun entries, viz. noise. Only 87 entries were useful proper noun entries, of which 74 entries or 22.4 percent of all the entries produced by the machine did not need any human editing.

O'Connor (40, 46) used for his study of the compatability of mechanized indexing with human indexing the existing retrieval system at a pharmaceutical research laboratory (Merck Sharp & Dohme Research Laboratories, West Point, Pa.). Several dozen documents were examined for each of the three terms, penicillin, toxicity, and mode of action. The only approaches considered were those involving occurrence and frequency of the term-word (e.g. toxicity) and synonymous and related words. For each indexing term, efforts were made to find a computer rule which would assign that particular term to just those documents assigned that term by the human indexers. It was established that, for instance, if "penicillin" was assigned as a term if the word "penicillin" occurred at least once, the result was overassignment up to ten percent of the entire document collection; if the term was assigned when the word occurred at least twice, the system would fail to assign "penicillin" to at least one tenth of the documents which should have had it. No general rules or conclusions were proposed.

A program to evaluate and compare the efficiency of machine indexing methods with human indexing with regard to the relevancy of documents retrieved was also reported by Donald J. Hillman (26), but neither

the results of the experiment nor a more detailed statement of the methods to be used in the evaluation are as yet available.

The effectiveness of machine indexing from titles has been evaluated on 1,500 technical titles, chiefly in the field of physics, by Baxendale (9). There were two criteria for evaluation. The index term had to be constituted solely of a noun and its adjective modifiers, and had to be meaningful with respect to the title. Using both of these criteria for evaluation, approximately 85 percent of the 1,500 titles were indexed with 100 percent accuracy. That is, all possible terms were selected and all satisfied both evaluation criteria. The effectiveness of the remaining 15 percent ranged between 95 percent and 40 percent accuracy.

A similar project comparing the results of automatic computer indexing of titles by the KWIC system with human indexing using a subject heading system was reported by Kraft (167). One source of data was 803 legal research projects and these titles indexed under a modified form of the Index of Legal Periodicals (ILP) system. The other source of data was 2,625 legal articles classified under the ILP system. Interpretation of data revealed, among other things, that 64.4% of the title entries contained as keywords one or more of the ILP subject heading words under which they were indexed; and 25.1% contained logical equivalents. The remaining 10.5%

of the title entries had nondescriptive titles. The author concluded that KWIC indexing of legal titles produces an index which costs less than a subject heading system in both time and cost of production and which ranks high in "findability."

In their study of automatic subject indexing from textual condensation, Slamecka and Zunde (52) examined a number of abstracts published in the Scientific and Technical Aerospace Reports by Documentation Inc. and compared their contents with the indexing terms which were assigned by human indexers to these documents. The results of the pilot experiment showed that, on the average, 80.4 percent of the index terms chosen by analysts were also contained in the abstract, and that each abstract contained an additional 10.9 terms which were part of the indexing vocabulary (Uniterm-type machine term vocabulary). It was also found that a condensation of approximately 83 percent was necessary in order to obtain significant indexing terms as the residue of a deletion process.

The above described investigations compared machine indexing by extraction with human indexing. Some other investigations were directed toward comparison of machine indexing by assignment (or automatic classification) with corresponding human performance.

In the attempt to measure the reliability of subject classification by men and machines as reported by Borko (12), three subject specialists

classified 997 abstracts for psychological reports into one of eleven categories. These abstracts were also mechanically classified by a computer program using a factor-score computational procedure. Each abstract was scored for all categories and assigned to the one with the highest score. The three manual classifications were compared with each other and with the mechanical classifications, and a series of contingency coefficients was computed. The average reliability of manual classification procedures was equal to .870. The correlation between automatic and manual classification was .766. Furthermore, it was concluded that humans will agree on the classification of approximately 75 percent of the documents, while automated classification procedures will agree with manual classification 59 percent of the time. Furthermore, by correcting the data for attenuation as a result of the known unreliability of the criterion, it was possible to determine that this percentage of agreement between automatic classification and perfectly reliable human classification could be raised to 67 percent.

Moreover the classes derived by factor analysis were compared with, and shown to be similar to, the existing subject classification system employed by the American Psychological Association. According to Borko, the study demonstrates the feasibility of using factor analysis as a method for determining the basic dimensions of a classification system.

Another evaluation experiment was carried out by Williams (61). Of 100 test documents initially selected, 17 were not completely indexed within the experimental structure. Therefore, complete results were available on only 83 of the original 100 documents. Sixty-three of these documents were classified by machine into only one category at both the major and minor levels of the predetermined hierarchial classification system. Twenty documents were classified into one category at the major level (higher generic level), and two categories at the minor level. When compared with the classification results by human indexers of the same documents, of the first group of documents 78 percent were correctly classified at the major level and 64 percent correctly classified at the minor levels. Of the second group (20 documents), 95 percent were correctly classified at the major level and 60 to 75 percent at the minor levels. According to Williams, two of the major reasons for misclassification were heterogeneous categories and small sample sizes. Since these results were obtained on only 15 reference documents per category, it is felt improvement could easily be achieved by increasing the number of reference documents.

In Stevens's (55) study, the number and type of descriptors assigned by machine were compared with those assigned by human indexers, both DDC and local. For the documents taken from the teaching sample, the average "hit" accuracy was 64.8 percent. For new or partially new input

(old items together with new) the "hit" accuracy or the percentage of descriptors originally assigned by DDC indexers which were also assigned by machine, was 48.2 percent. No significant difference in the average accuracies was obtained as between using titles-and-cited-titles only and using titles-and-abstracts from the same items. In another evaluation effort, 25 of the items in the test runs were submitted to one or more members of the NBS staff, all of whom were users of the collection. They were asked to choose 12 descriptors for each item exclusively from the list of descriptors actually available to the machine. The percentage of identical descriptors thus chosen were from 40 percent to 54.2 percent. Thus the results appear to fall within the range of agreement-data for human interindexer consistency.

1.3.2. Comparing Various Machine Indexing Methods

As yet, only a limited amount of research has been done to compare one machine indexing method with another or how various machine indexing methods perform on the same input and for the same user requirements. For machine indexing by extraction Baxendale (7) compares subject indexes produced by simple deletion of non-significant terms, by selection by topic sentences and deletion, and by selection by prepositional phrases and deletion. She arrives at the conclusion that high percentages of condensation are possible by all of the techniques outlined without untoward loss of content of an article. No clear advantage of one of these methods

against the other in selection of indexing terms was demonstrated, except that selection by prepositional phrases enables the system to produce precoordinated terms, which under certain conditions might be preferable to Uniterms.

Borko and Berwick (14, 15) made a comparative study of two methods of indexing by assignment (automatic classification). To test the hypothesis that the classification system derived by factor analysis provides a sound basis for document classification and is compatible with other systems, the same corpus of documents was selected as used by Maron in his automatic indexing experiment. The following procedural steps for automatically classifying the documents were used. First, each document, in machine readable form, was analyzed by the computer. A list of the index terms and their frequencies of occurrence in each document was recorded. Second, the category, or categories, containing the index term was assigned a value equal to the product of the number of occurrences of the word in the abstract and the normalized factor loading of the word in the category. If more than one index term appeared in a category, the products were summed. Thus

$$P = f(L_1 \times T_1 + L_2 \times T_2 + ... + L_n \times T_n)$$

where

P = predicted classification, L_n = normalized factor loading of term n for a given category, and T_n = number of occurrences of the n-th term.

Third, after each index term had been considered, the category having the highest numerical value was selected as the most probable subject classification for the document in question. Of the 90 documents in the validation group which contained two or more cue words, and which therefore could be automatically classified, 44 documents, or 48.9 percent, were placed into their correct categories by use of a computer formula. These results were almost identical to those obtained by Maron in a previous experiment using the same data but with a different set of classification categories and a different computational formula. In classifying the documents in the experimental group Maron's technique was, however, superior. There the percentage of correctly classified documents was 84.6% by Maron as against 63.4% by Borko. Obviously, the factor technique did poorly when operating on the specific body of data on which the classification system and the factor loadings were derived. A possible explanation is that the factor analysis method is a generalizing technique designed to deal with common properties and not with the specific variances found in a population sample. In contrast, Maron's technique capitalizes on the specific variance in the sample and, therefore, did far better in the automatic classification of the documents in the experimental group than for the validation group. Consequently, for Maron's technique, the statement that "the more cue words in the document, the better the automatic indexing" applies. In contrast, a prediction technique based upon factor loadings appears to have little dependence on the number of cue

words in the article. That is to say, the number of documents containing one or two cue words were classified with almost the same degree of accuracy as those containing a larger number. This makes sense when one realizes that factor analysis is a generalizing technique designed to minimize the specific variance of the individual words. As a result, a method of automatic document classification based upon factor loadings enables one to classify documents containing a minimum of index terms.

However, since the nature of that study did not provide for an isolation of the techniques used in automatic classification from the categories themselves, a new series of tests were conducted. Three hypotheses were tested. They were: (1) using the original classification schedule, automatic document classification will be more successfully performed by means of a Bayesian prediction equation (Maron's method) than by factor scores; (2) using the modified classification schedule, automatic document classification will be more successfully performed by means of a Bayesian prediction equation than by factor scores; and (3) documents will be correctly classified in the modified classification schedule in a number significantly greater than in the derived classification scheme using either the Bayesian or the factor score procedures for automatic document classification.

It was concluded that there was no statistically significant difference in the ability of these two procedures to automatically classify

documents. The comparison of the effectiveness of the original and the modified classification categories for automatic document classification proved that more documents were correctly classified when using the modified schedule than by using the original and that the increase was a statistically significant one in the most important case when predicting the classification of the previously unexamined documents in the validation group. Borko and Bernick, therefore arrived at the following three conclusions. First, it is possible to mathematically derive a set of classification categories that are descriptive of the major content dimensions of a population of documents. Furthermore, these dimensions are relatively stable as long as the parent population is itself stable and unchanging. Second, automatic document classification is possible and may be accomplished by use of either Bayesian or factor score procedures. Third, if automatic document classification is to be used, superior results will be obtained by using mathematically derived classification categories based upon statistical analysis of the words in the documents and statistical indexing techniques.

In the opinion of the authors, factor analysis has been demonstrated to be a useful technique for determining the major dimensions in an unstructured mass of material. It has been used to derive classification categories for computer literature and for psychological reports. In both cases the classification categories were reasonable and reliable. Factor analysis can be applied to unstructured subject matter such as

newspaper articles, intelligence reports, etc., in an attempt to derive a reasonable and useful set of classification categories for this type of material.

I.4. TIME AND COST ANALYSES

Little has been reported on the time studies of processing machine generated indexes and even less on the costs of automatically creating indexes. Short references to processing time are in the Kraft (28) and Levery (30) papers only. Kraft reports that using an IBM 1401 computer with 8,000 memory positions and tape drives, the following processing times were observed for the auto-indexing run:

(a) using an exclusion list of 600 common words---16 seconds per document.

(b) using an exclusion list of 600 common words and an accept list of 2,200 words---60 seconds per document.

The above times include card reading, auto-indexing, and tape writing.

Levery (30) reports processing time of 15 seconds per document. It is not specified in the report whether this includes all the time for matching terms, calculating their absolute and relative frequencies, etc., but it may be assumed that it does.

Artandi presented cost estimates for mechanical book indexing and for mechanical indexing of proper nouns. For book indexing by the method of matching the text against a vocabulary on the IBM 1620 computer, Artandi

(6) quotes $2.046 per page with an initial investment from $2.71 to $0.014 per page for 5 to 1,000 books over a period of 10 years. The first or operating costs are broken down into the following items:

conversion of text for machine input	$ .712
machine time and labor for one run	.979
alphabetization, elimination of duplicates, crossreferences, material (est.)	.355
	$ 2.046

For mechanical indexing of proper nouns on the IBM 1620 computer, Artandi (4) quotes $2.06 per text page for 5 books and $1.92 per text page for 100 books as compared with $0.04 per text page if indexing is done by conventional methods, viz. by human indexers.

1.5. CONCLUSIONS AND RECOMMENDATIONS

Studies and experiments, which have been done on automatic indexing, indicate that operational systems of this type are entirely possible in principle. No pioneering discoveries are required to have the machine read the document and index it either by extracting pertinent terms from the text of the document or by assigning terms based on the document analysis. However, a considerable amount of research is still required in order to have the machine do it well and efficiently. Thus, the problem is basically that of optimization: optimizing index file structure and organization, improving term selection criteria, applying methods of linguistic analysis for class identification, etc. The problem is also one of cost: under what circumstances does it pay to have indexing done by machine.

A large amount of statistical analysis is needed to establish significance criteria for selecting words as indexing terms by the frequency of their occurrence. The relative-frequency concept of word significance should be compared with the simple-frequency approach. Word-frequency counts for specialized subject-fields need to be conducted and utilized for establishing profile parameters. Functions that derive a measure of significance from the relative frequency of a word should be compared for ease of interpretation and computation, and for amount of discrimination. Statistical criteria for selecting words (Uniterms) and precoordinated terms should be devised,

and the merits of these methods should be carefully evaluated. Studies should also be made of the optimal number of terms per document, whether the number of terms are related to the number of words in the document and the total number of documents in store and how they are so related, and whether it is desirable or possible to predetermine the ratio of single words to pre-coordinated terms in the index, if both types of terms are used.

Additional studies are needed to investigate the effectiveness of non-statistical measures of significance, such as positional or pragmatic measures. It should be established whether it is necessary and/or desirable to delete common words by using a stop list when indexing is done by statistical method (frequency count) and it should be determined what the optimal size of such a stop list should be. The effect on index quality of matching terms on a predetermined number of characters should be studied as well.

If the quantitative procedures combined with the simple non-probabilistic measures of word significance do not produce the desired refinement of the indexing system, consideration might be given to qualitative analysis, such as investigating synthetic or linguistic relations. However, quantitative methods should be preferred over qualitative ones which require interpretation of the text because quantitative methods are much less complicated and, therefore, much less costly and time consuming.

For indexing by assignment or automatic categorization, the way to improve the systems efficiency is by conducting more extensive studies on the relations of words or word combinations and categories of various classification systems, the degree of the resolution of the classification system, definitions of class profiles, and word significance coefficients. The size of the sample of documents for the determination of the total population parameters and the number of the terms assigned to the documents as well as their generic relations should also be analyzed.

It might be advantageous to combine both the indexing by extraction and the indexing by assignment methods in one system. This might provide for better term selection and assignment control possibilities.

In all the systems proposed thus far, one element is generally missing, the absence of which hardly justifies calling these systems automatic. This missing element is the feedback loop, which is essential for any automatic system expected to react to changing input conditions. Schematically, the present systems under study can be represented by the following diagram (Fig. 1).

Fig. 1

For automatically operating systems, there should be at least one feedback loop from the indexing output (Fig. 2).

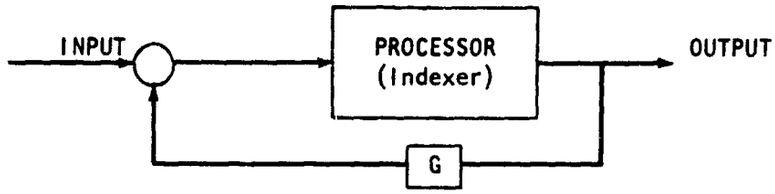

Fig. 2

A more advanced system should have another feedback loop from the retrieval output to the input into the system: (Fig. 3)

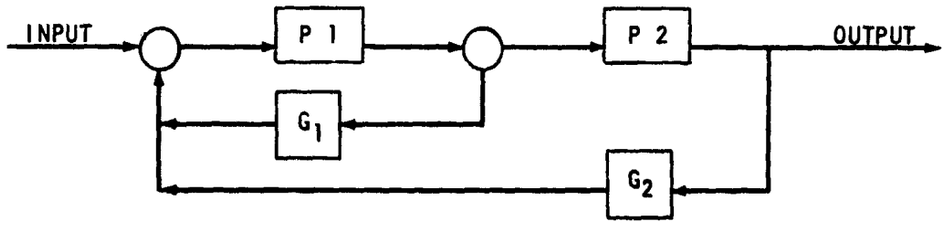

Fig. 3

The two feedback loops are necessary to adjust indexing parameters according to the changing quality and quantity of input material and user's requirements and to control the index file organization for high efficiency of operation. This is especially important for systems processing large amounts of data. The effects of file organization on systems efficiency and other related optimization problems were recently investigated by Zunde (281).

PART II

FORMAL AUTOINDEXING OF SCIENTIFIC TEXTS (FAST)

FEASIBILITY AND SYSTEMS STUDY

II.1. CHARACTERISTICS OF A SCIENTIFIC UNITERM INDEX

The automatic indexing system, which was to be designed under this contract, had to replace human indexing of scientific abstracts in projects such as Interagency Life Sciences Supporting Space Research and Technology Exchange (ILSE) of the Department of Defense or NASA or in similar projects where short scientific texts, available in machine readable form, are to be indexed for retrieval. Samples of such abstracts are shown in Annex 1. It was required that the documents would be indexed by the Uniterm (coordinate) method as it was the case when documents were indexed by humans. The information was to be stored on magnetic tapes and searches were to be made by the computer.

Prior to the design of a mechanized substitute for human indexing, for this type of input material, characteristic features and parameters of a typical index produced by humans were investigated. The ILSE Index to the store of research abstracts for the year 1963 was selected as a characteristic sample. The total number of indexed documents in store was 2,809. The system's vocabulary contained 3,146 indexing terms. Since some of these indexing terms were hyphenated, the actual number of Uniterms or single words was 3,210. The total number of postings was 37,471, so that on the average there were 11.91 postings per indexing term and 13.34 postings per document. Since the research tasks were basically oriented toward life sciences, life science terminology prevailed to a certain extent in the vocabulary, but

generally the terms were not too specific. Some sample pages of the ILSE 1963 Index vocabulary are reproduced in Annex II.

The population of 3,210 single words (Uniterms) which appear in the index were analyzed structually. For each word, the number of syllables was determined and counted. The results are shown in Table 1.

Table 1. Breakdown of Uniterms by the number of syllables in the ILSE 1963 Uniterm vocabulary.

No. of Syllables (i)	No. of Words with i Syllables [T(i)]	Relative Frequency [t(i)]	Total No. of Syllables in the Category [i · T(i)]
1	429	0.1337	429
2	758	0.2363	1,516
3	767	0.2391	2,301
4	635	0.1979	2,540
5	357	0.1110	1,780
6	181	0.0562	1,085
7	59	0.0184	419
8	21	0.0065	168
9	2	0.0006	18
10	1	0.0003	10
TOTAL	3.210	1.0000	10,266

Furthermore, for the same population of words, the number of letters was counted for each syllable and thus the frequency of syllables of various lengths was obtained (see Table 2).

The following average values are readily obtained from the above data:

Average number of syllables per word	$\bar{i} = 3.2001$
Average number of letters per syllable	$\bar{x} = 2.7024$
Average number of letters per word	$\bar{t} = 8.6480$

Table 2. Breakdown of syllables by number of letters for the Uniterms in the ILSE 1963 index vocabulary.

No. of Letters Per Syllable (x)	No. of Syllables With j Letters [R (x)]	Relative Frequency [(x)]	Total No. of Letters in the Category [x R(x)]
1	1,200	0.1169	1,200
2	3,411	0.3352	6,822
3	3,444	0.3353	10,332
4	1,750	0.1704	7,000
5	384	0.0374	1,920
6	72	0.0070	432
7	3	0.0003	21
8	2	0.0002	16
TOTAL	10,266	1.0000	27,743

Independently of the above counts by syllables, a character count was made for each of the 3,146 indexing terms (the hyphen in hyphenated terms, such as in MAN-MACHINE, was this time counted as a character). Table 3 shows a breakdown of the indexing terms by number of characters and a plot of the corresponding frequency distribution is given in Figure 1.

From Table 3 we find that the average number of characters for indexing term is $\bar{t}' = 8.899$.

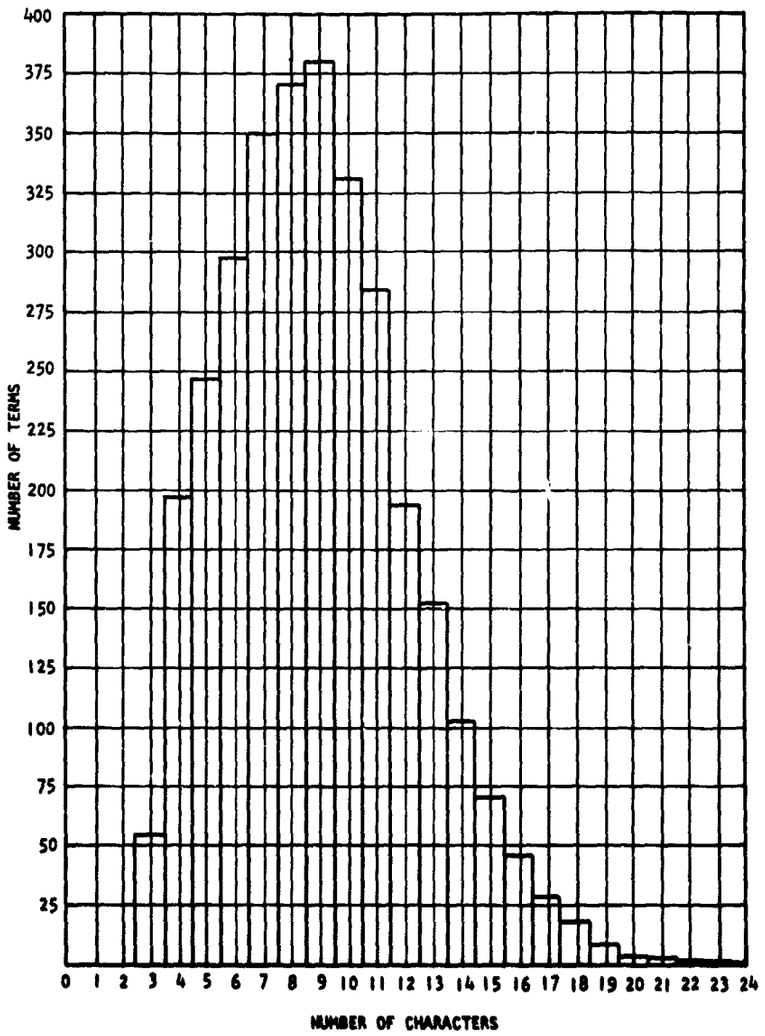

Figure 1 - Frequency distribution of indexing terms by the number of characters.

Table 3. Breakdown of indexing terms by number of letters (characters) of the ILSE 1963 vocabulary.

No. of Characters (t)	No. of Terms M(t)	Total No. of Characters in the Category t • M(t)
3	55	165
4	199	796
5	246	1,230
6	299	1,794
7	350	2,450
8	373	2,984
9	379	3,411
10	329	3,240
11	285	3,135
12	194	2,328
13	151	1,963
14	102	1,428
15	71	1,065
16	46	736
17	29	493
18	17	306
19	9	171
20	4	80
21	4	84
22	3	66
24	1	24
TOTAL	3,146	27,999

Figure 2 gives the distribution of subject word lengths by number of characters for the Stanford Research Institute Uniterm dictionary, containing 2,082 single word descriptors,*) and Figure 3 gives the distribution of the

*) The plot was reproduced from a paper by Ch. P. Bourne and D. F. Ford on the statistics of letters in English words (84).

word lengths of 5,153 most frequent words selected from a sample of popular magazines.**) The corresponding distribution of terms in the ILSE 1963 vocabulary is shown for the purpose of comparison.

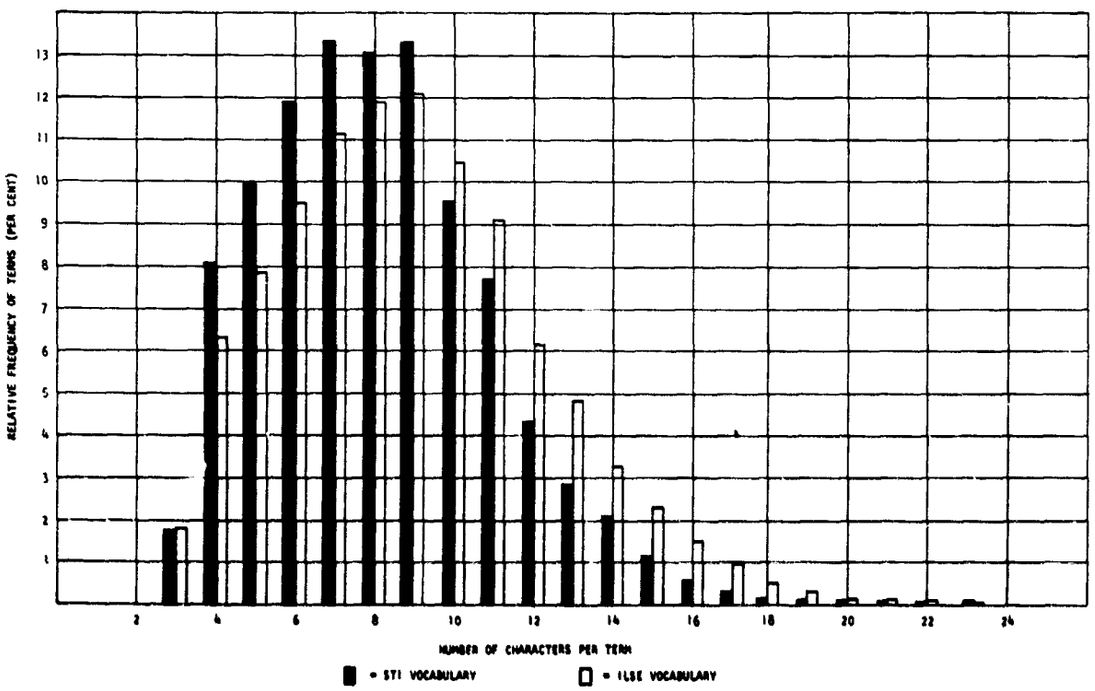

Figure 2 - Distribution of subject word length f(t) in the SRI and ILSE vocabularies.

By comparing the plots on the Figure 2 and Figure 3, one can see that the distribution of subject word lengths in the SRI Uniterm vocabulary and ILSE 1963 Uniterm vocabulary is rather similar, whereas the distribution of

**) The data was taken from a paper by E. S. Schwartz (231).

word lengths of most frequent words in a sample of popular magazines is significantly different from both distributions in SRI and ILSE vocabularies.

We shall investigate next, whether the differences in the distribution of word lengths in natural language (which is represented by the sample of popular magazines) and in "indexing language" such as the SRI and ILSE vocabularies of scientific terms are basically due to different parameters (average length of terms and variance) of one and the same distribution function, or whether the differences result from the existence of entirely different distribution laws for these families of words.

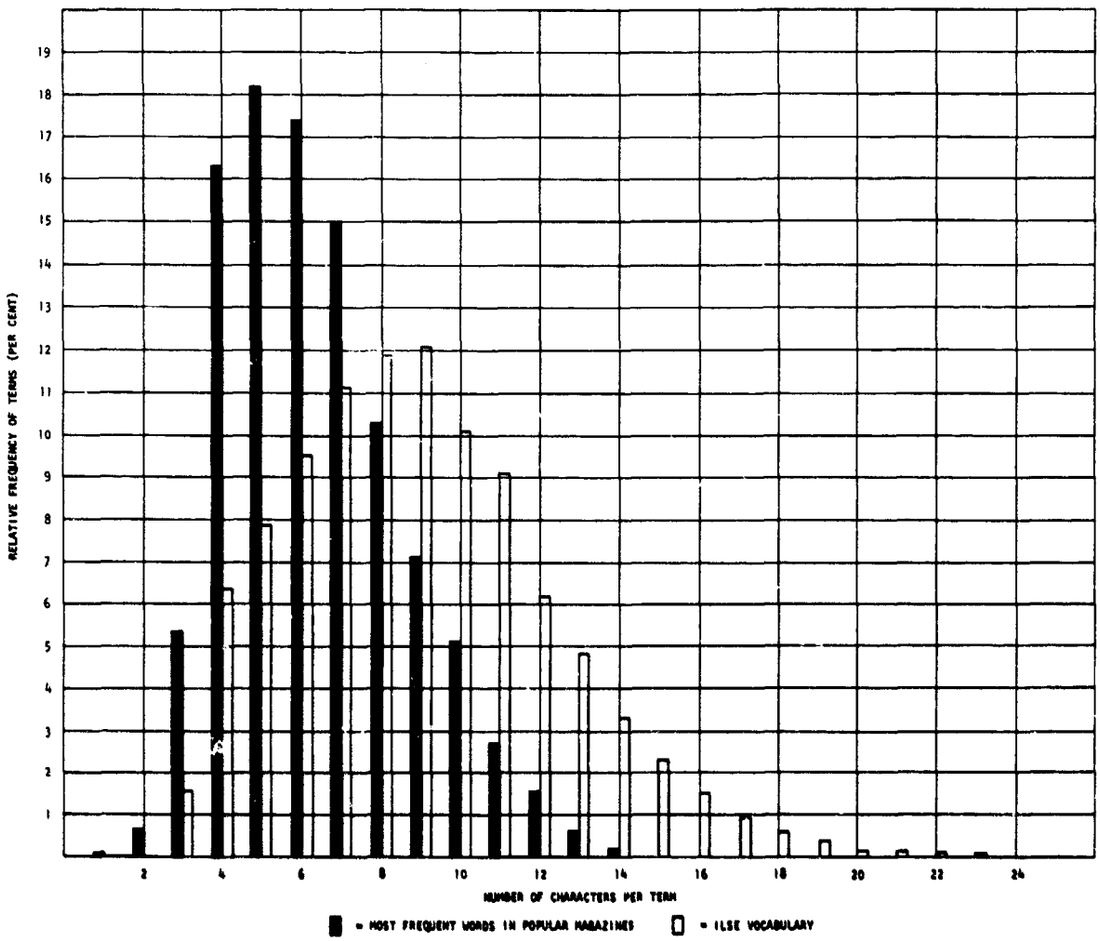

Figure 3 - Distribution of word lengths f(t) of 5,153 most frequent words in a sample of popular magazines and the distribution of subject word lengths in ILSE 1963 vocabulary.

II.2. FORMATION OF WORDS IN THE INDEXING LANGUAGE

W. Fucks proposed in his paper (119) a mathematical model of the word formation out of syllables and syllable formation of letters in natural language texts. Based on the investigations of the process of the formation of words out of syllables, he derived the following theoretical probability distribution function

$$p(i) = \frac{e^{-(i-1)}\,(\bar{i}-1)^{i-1}}{(i-1)\,!} \qquad (1)$$

where $p(i)$ is the probability of occurrence of words with i syllables, $i = 1,2,3,\ldots n$, and $\bar{i}$ - average number of syllables per word.

For the probability $v(x)$ of a syllable having x letters, $x = 1,2,3,\ldots m$, the following modified equation was obtained

$$v(x) = e^{-(\bar{x} - \sum_{1}^{\infty} \varepsilon_K)} \sum_{\nu=0}^{\infty} (\varepsilon_\nu - \varepsilon_{\nu+1}) \frac{(\bar{x} - \sum_{1}^{\infty} \varepsilon_K)^{x-\nu}}{(x-\nu)\,!} \qquad (2)$$

where ε_K, $K = 0,1,2,3,\ldots l$ are special parameters of a given linguistic structure.

Now, does the above formula represent a valid law of the fundamental properties of the word formation process in the "indexing language" as well, or do the "indexing languages" obey laws of their own?

The relative frequency distribution $p(i)$ of syllables per indexing term or Uniterm ($i = 1,2,3,\ldots m$ syllables) for the scientific ILSE Uniterm vocabulary is shown in Figure 4. In the same illustration, there

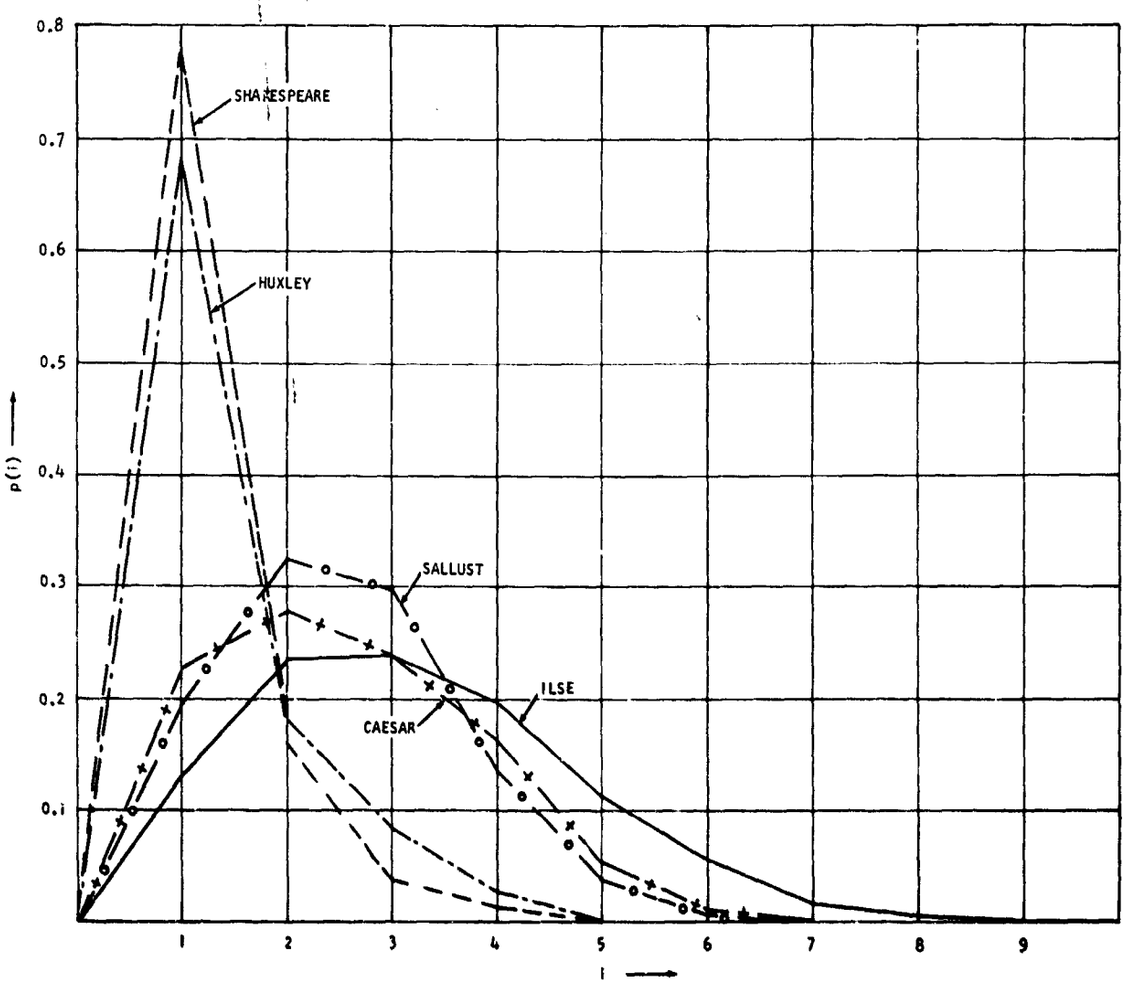

Figure 4: Relative frequency distribution p(i) of syllables per word in four different texts and in the ILSE 1963 Uniterm vocabulary (i = number of syllables = 1, 2, 3, ...).

have been plotted relative frequency distributions p(i) of words in Shakespeare's Othello, Huxley's Antic Hay, as well as two curves derived from Latin texts, i.e., Sallust's Bellum Jugurthinum, and Caesar's De Bello Gallico. Table 4 gives the values of the mean $\bar{i}$, variance σ,

entropy S, skewness ρ_3, kurtosis ρ_4, and third and fourth order moments, μ_3 and μ_4, with regard to the mean value for these five population samples. Fucks called these values style characteristics, since they characterize the style of a particular author. We see that the Uniterm set of indexing terms is different in character from either of the four other samples.

Table 4: Style characteristics of the ILSE Uniterm Vocabulary and the two English and two Latin texts from Figure 4.

SOURCE	$\bar{i}$	σ	μ_3	ρ_3	μ_4	ρ_4	S
ILSE Uniterm	3.2001	1.5289	2.2464	0.6286	27.4959	4.0327	0.7779
Shakespeare	1.2758	0.5954	0.5040	2.3875	1.1206	8.9149	0.2883
Huxley	1.4087	0.7770	0.7745	1.6510	2.0859	5.7226	0.3804
Sallust	2.5102	1.1059	0.6377	0.4715	4.2977	2.8732	0.6405
Caesar	2.5368	1.2234	0.9097	0.4970	5.8172	2.5971	0.6719

A comparison was also made of the values of the distribution function p(i) for the "indexing language" and the average values for nine languages derived from many texts of many authors (see Table 5). The latter figures are taken from the already quoted paper of W. Fucks (119).

Table 5. Relative frequency distributions, mean values and entropies of the ILSE 1963 Uniterm vocabulary and of nine languages taken from a representative average of texts (syllables per word).

	UNITERM	ENGLISH	GERMAN	ESPERANTO	ARABIC	GREEK	JAPANESE	RUSSIAN	LATIN	TURKISH
p(1)	0.1337	0.7152	0.5560	0.4040	0.2270	0.3760	0.3620	0.3390	0.2420	0.1880
p(2)	0.2363	0.1940	0.3080	0.3610	0.4970	0.3210	0.3440	0.3030	0.3210	0.3784
p(3)	0.2391	0.0680	0.0938	0.1770	0.2239	0.1680	0.1780	0.2140	0.2870	0.2704
p(4)	0.1979	0.0160	0.0335	0.0476	0.0506	0.0889	0.0868	0.0975	0.1168	0.1208
p(5)	0.1110	0.0056	0.0071	0.0082	0.0017	0.0346	0.0232	0.0358	0.0282	0.0360
p(6)	0.0561	0.0012	0.0014	0.0011		0.0083	0.0124	0.0101	0.0055	0.0056
p(7)	0.0184		0.0002			0.0007	0.0040	0.0015	0.0007	0.0004
p(8)	0.0065		0.0001				0.0004	0.0003	0.0002	0.0004
p(9)	0.0006						0.0004			
p(10)	0.0003									
$\bar{i}$	3.2001	1.351	1.634	1.859	2.104	2.105	2.137	2.228	2.392	2.455
S	0.7779	0.367	0.456	0.535	0.513	0.611	0.622	0.647	0.631	0.629

It can be seen from Table 5 that the ILSE "indexing language" is not similar to or identical with any of the above languages either. Incidentally, the average number of syllables per term is much closer to the average in Turkish texts than in English.

With $\bar{i}$ = 3.2001 for ILSE indexing terms, the theoretical or expected distribution, calculated from the Eq. (1), is given in column 1 of the Table 6. Column 2 of that table gives the actual distribution. These distributions are also plotted in Figure 5.

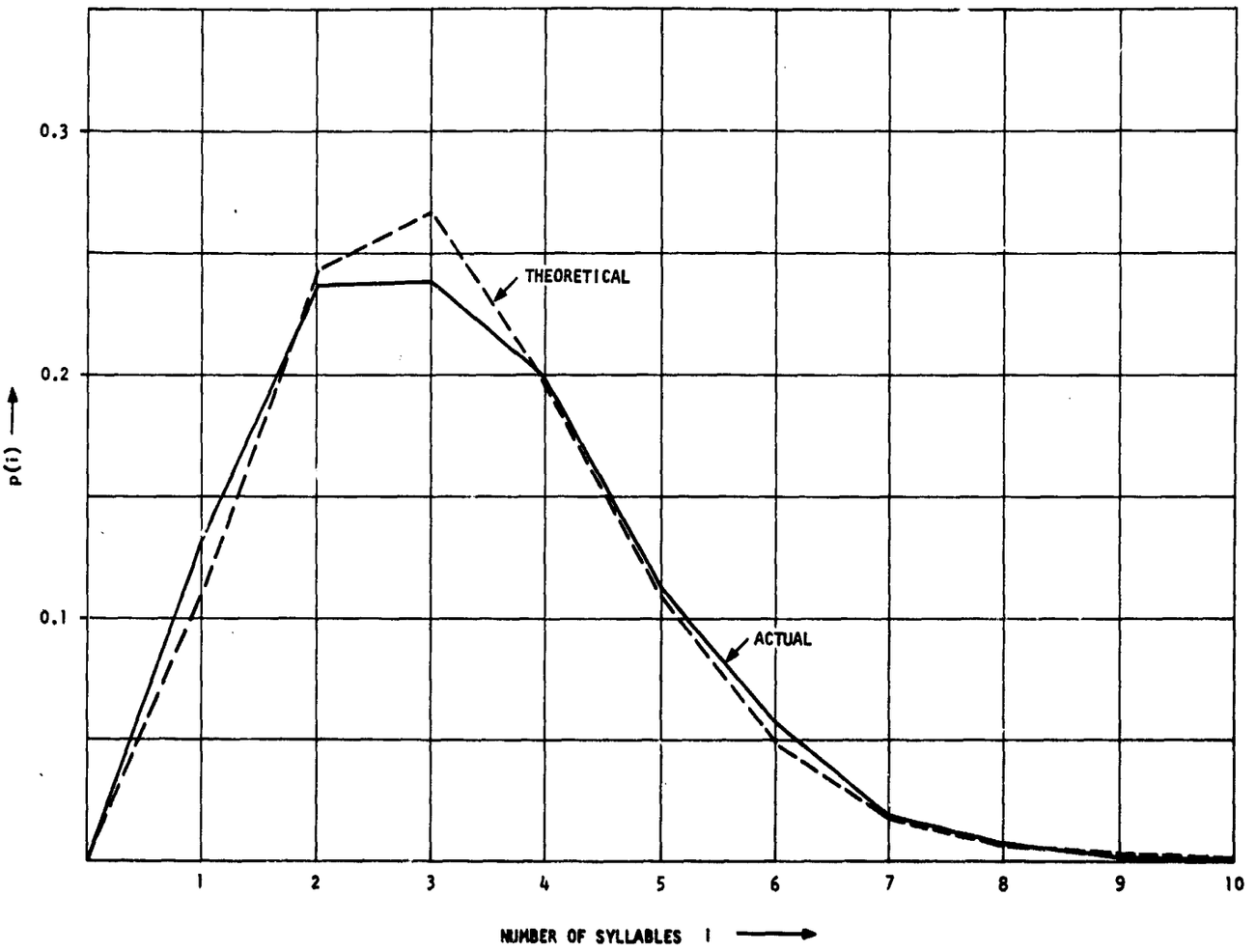

Figure 5: Relative frequency distribution $p_{(i)}$, theoretical and actual, of the ILSE 1963 vocabulary Uniterms (syllables per word).

The comparison of the two curves, which represent the actual distribution and the expected distribution as calculated from Eq. (1), shows that they agree fairly well and that Eq.(1) reflects at least the main features of the process of formation of words out of syllables for the "indexing language" as

it does for the natural languages. It remains now to investigate how finer characteristics of the word formation in the "indexing language" can be derived.

Table 6. Theoretical and actual frequency distribution of indexing terms in ILSE 1963 Uniterm vocabulary by number of syllables (i).

p(i)	Relative Frequency	
	Theoretical	Actual
p(1)	0.1109	0.1337
p(2)	0.2435	0.2363
p(3)	0.2679	0.2391
p(4)	0.1966	0.1979
p(5)	0.1083	0.1110
p(6)	0.0477	0.0561
p(7)	0.0176	0.0184
p(8)	0.0057	0.0065
p(9)	0.0015	0.0007
p(10)	0.0003	0.0003

To obtain the distribution of the number of letters in syllables, we shall use the Eq. (2). The parameters ε for that equation are found as follows.

First we derive the characteristic function for that distribution. It appears to be:

$$M(ju) = \sum_{x=0}^{\infty} v(x)\, e^{jux}$$

$$= e^{-(\bar{x}-\sum_{1}^{\infty}\varepsilon_k)} \sum_{x=\nu}^{\infty} \sum_{\nu=0}^{\infty} (\varepsilon_\nu - \varepsilon_{\nu+1}) \frac{(\bar{x}-\sum_{1}^{\infty}\varepsilon_k)^{x-\nu}}{(x-\nu)!}\, e^{jux}$$

$$= e^{(\bar{x}-\sum_{1}^{\infty}\varepsilon_k)(e^{ju}-1)} \sum_{\nu=0}^{\infty} (\varepsilon_\nu - \varepsilon_{\nu+1})\, e^{ju\nu} \qquad (3)$$

From the characteristic function we can derive moments of any order by noting that:

$$M_n = \lim_{u \to 0} \frac{1}{j^n} \frac{d^n M}{d u^n} = \lim_{u \to 0} \frac{1}{j^n} \frac{d^n}{d u^n} \left[\sum_{x=0}^{\infty} v(x) e^{jux} \right] \qquad (4)$$

$$= \sum_{x=0}^{\infty} x^n v(x)$$

Where M_n is the n-th order moment about the origin.

Assuming that $\varepsilon_0 = \varepsilon_1 = 1$, $\varepsilon_4 = \varepsilon_5 = \ldots\ldots = 0$ and $\varepsilon_2 \neq 0$, $\varepsilon_3 \neq 0$, we get

$$\mu_1 = (\bar{x} - \sum_{1}^{\infty} \varepsilon_K) + 1 + \varepsilon_2 + \varepsilon_3 = \bar{x} \qquad (5)$$

$$\mu_2 = \bar{x}^2 + \bar{x} - 1 - (\varepsilon_2 + \varepsilon_3)^2 + 2\varepsilon_2 \qquad (6)$$

$$\mu_3 = \bar{x}^3 + 3\bar{x}^2 - 2\bar{x} - 3\bar{x}(\varepsilon_2 + \varepsilon_3)^2 + 6\bar{x}\varepsilon_3 \qquad (7)$$

$$- 3(1 + \varepsilon_2 + \varepsilon_3)^2 + 2(1 + \varepsilon_2 + \varepsilon_3)^3 - 6(\varepsilon_2 + \varepsilon_3)(\varepsilon_2 + 2\varepsilon_3) + 6\varepsilon_3$$

With regard to the mean, the corresponding moments are:

$$m_1 = 0 \tag{8}$$

$$m_2 = \bar{x} - 1 - (\varepsilon_2 + \varepsilon_3)^2 + 2\varepsilon_3 \tag{9}$$

$$m_3 = \bar{x} - 3(1 + \varepsilon_2 + \varepsilon_3)^2 + 2(1 + \varepsilon_2 + \varepsilon_3)^3 - 6(\varepsilon_2 + \varepsilon_3)(\varepsilon_2 + 2\varepsilon_3) + 6\varepsilon_3 \tag{10}$$

From our sample population we have:

$$\bar{x} = \bar{l} = 2.7024$$
$$m_2 = \mu_2 - \mu_1^2 = 1.1167$$
$$m_3 = \mu_3 - 3\mu_2\mu_1 + 2\mu_1^3 = 0.3717$$

With

$$\varepsilon_2 = -\varepsilon_3 \pm \sqrt{-1 + \bar{x} + 2\varepsilon_3 - m_2} = -\varepsilon_3 \pm \sqrt{2\varepsilon_3 + 0.5857}$$

From Eq. (8), we obtain by substitution in Eq. (9)

$$8\varepsilon_3^3 - 7.0284\,\varepsilon_3^2 + 0.6219 = 0$$

Hence

$$\varepsilon_3 = 0.405$$

and

$$\varepsilon_2 = 0.777$$

Substituting these values into Eq. (2), we get the following distribution function of letters in syllables for Uniterm indexing terms:

$$v(x) = 0.5943 \left[0.223 \frac{0.5204^{x-1}}{(x-1)!} + 0.372 \frac{0.5204^{x-2}}{(x-2)!} + 0.405 \frac{0.5204^{x-3}}{(x-3)!} \right]$$

Table 7 gives theoretical distribution and actual distribution of letters per syllable and Figure 6 shows the plot of these two curves. There again is a satisfactory correspondence between theoretical and actual values and better fits could be obtained by introducing additional parameters ϵ, calculated from higher order moments.

Table 7: Theoretical and actual frequency distribution of the number of letters in syllables in the ILSE 1963 Uniterm vocabulary.

No. of Letters Per Syllable (x)	Relative Frequency	
	Theoretical	Actual
1	0.1325	0.1169
2	0.3590	0.3352
3	0.3745	0.3355
4	0.1583	0.1705
5	0.0381	0.0374
6	0.0064	0.0070
7	0.0012	0.0003
8	0.0001	0.0002

Thus we can conclude that certain probabilistic laws do govern the formation of indexing terms from syllables and letters. If the "style characteristics" viz. moments of various orders of the distribution of terms by syllables and letters are known or can be obtained from

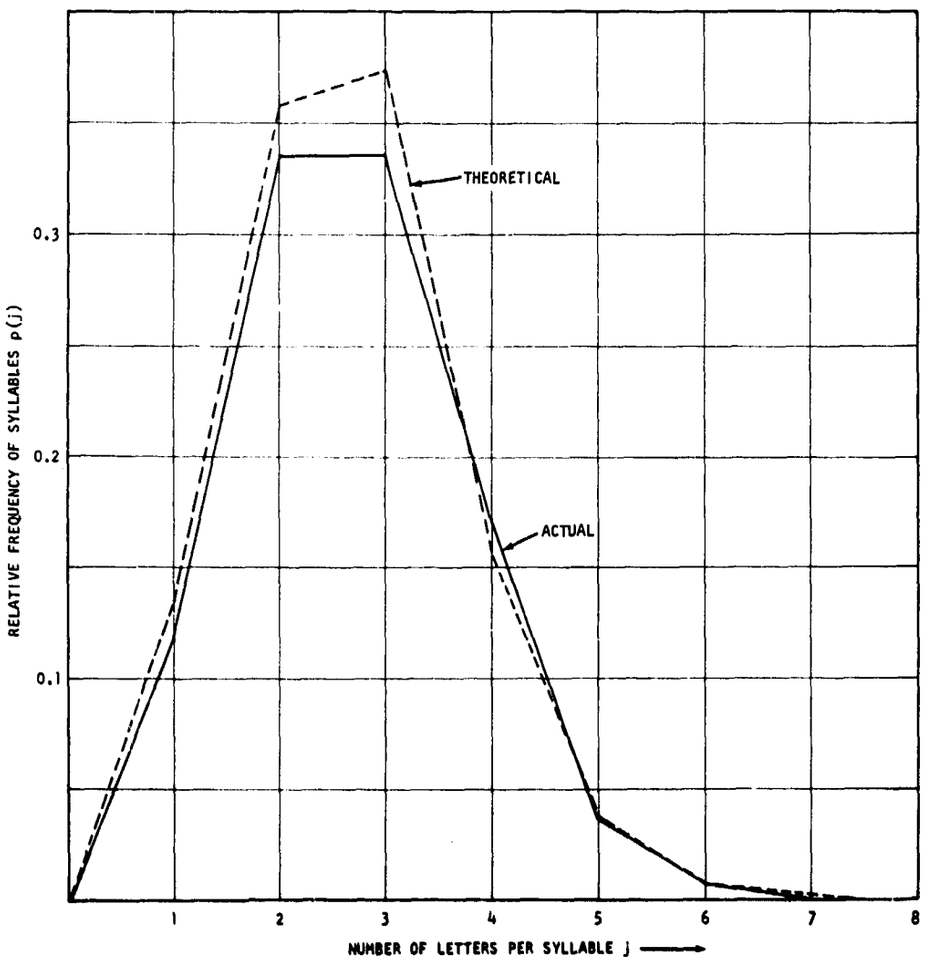

Figure 6: Relative frequency distribution, theoretical and actual, of syllables by the number of letters in ILSE 1963 Uniterm vocabulary.

a representative sample, it should be possible to theoretically calculate the most probable distribution of the terms for populations of any type and size with satisfactory accuracy.

This is of great practical value in calculating required memory space to store lists of terms in computer memory, designing match procedures, deriving the number of significant characters for terms on the authority lists, and optimizing systems performance. Applications of this kind were made in designing the Formal Autoindexing of Scientific Texts (FAST) System described in the following chapters.

11.3 FORMAL AUTOINDEXING OF SCIENTIFIC TEXTS (FAST) SYSTEM

It is assumed that the input into the system, which consists of short scientific abstracts, is available in computer readable form. At this stage of development, this means that the text is available on magnetic tapes, although the method of reading the abstracts by the machine is immaterial to the FAST program. For instance, magnetic tapes could be replaced by optical scanning devices, in which case the abstracts would be read from printed copies.

There are no particular requirements for the conversion of the texts of the abstracts to machine readable form except that the words should not be broken apart at the end of the line for the purpose of carrying them over. However, it is possible that certain requirements might be originated by the user as part of the overall systems specifications, for instance, fixed positions for certain proper names, spelling of chemical compounds, etc.

The indexing terms are extracted from the abstracts as the computer scans the text word by word. Blank spaces indicate to the computer the beginning and the end of a word. The essential parts of the FAST system are: a programmed mechanism for eliminating words which under no circumstances can be considered potential indexing terms (Kill List Program), a programmed mechanism for selecting, editing and cumulating significant terms (Authority File Program) and a programmed mechanism for implementing human control and optimization capability in unresolved cases (Residue Editing Program).

The mechanism for eliminating words which are unacceptable indexing terms consists of a set of computer instructions to delete terms which match with the terms on the Kill List especially designed for this system. The match has to be complete on all characters for the computer to delete the word.

Every word which is put into the system, be it title, body or footnote of the abstract, is matched against the Kill List, but there is one exception to the deletion instruction. Certain words, though they appear on the Kill List, are not deleted if they are part of the title. Therefore, before deleting the words in the titles, the computer compares them with a Title Exemption File. If a word appears in that file, it is not deleted as would be the case if it were found in the body of the abstract, but is flagged and retained as indexing term.

The reason for this is that there is a category of words, which under most circumstances would be undesirable indexing terms, but in certain cases might become acceptable. Consider words such as ATTENTION, DURATION, OPINION, WORK, etc. In sentences like: "The investigator paid much attention to the proper selection of test animals" or "The work progressed satisfactorily," the words attention and work would not be significant enough to justify their selection as indexing terms. But let's take now the sentences: "Investigation of the factors influencing the attention of astronaut under severe flight conditions" or "Measuring the efficiency of work of primates." There the same words attention and work are significant indicators of the content of the documents to be indexed. It has been established that usually words of this

type become significant indexing terms if the processes or objects they designate are subjects of a study or investigation. In such cases there is a high probability that these words will appear in the titles or headings of the abstracts describing such scientific tasks or projects. The function of the above described exemption mechanism for titles is to detect such words and convert them to indexing terms.

From what remains after deletion of insignificant words, the computer selects, edits, and cumulates significant indexing terms (Authority File Program). The basic element of this mechanism is a file of terms considered to be acceptable indexing terms for the particular type of input. This file is called the Authority List. For different subjects fields of input, Authority Lists might be different.

Words in the abstract are matched against the terms on the Authority List. However, this time a complete match is not required on all characters but only on certain significant characters which are specifically identified for each term on the Authority List (see Annex III). The longest match, if there is a match at all, of a given word from the text on the significant characters of a term in the Authority List is considered a "hit." If there is a "hit," the term on the Authority List is accepted and printed as the indexing term.

To illustrate the procedure, consider the word CONDITIONAL. The Authority List might contain terms (asterisk indicates the end of significant characters) such as

CONDITION*	(9 significant characters)
CONDITIONE*d	(10 significant characters)
CONDITIONI*ng	(10 significant characters)

The FAST program will start matching the word CONDITIONAL agaisnt CONDITIONING and then against CONDITIONED, since these two have the greatest number of significant characters in the batch of terms against which the word CONDITIONAL is matched. Since the word does not match with either of these terms on 10 significant characters, it is next matched against the term CONDITION which requires matching on 9 significant characters. The word CONDITIONAL does match on the first nine characters of the term CONDITION on the Authority List, and therefore, the term CONDITION (but not the word CONDITIONAL) is assigned to the corresponding abstract as indexing term.

The subsets of terms of the Authority List, against which a word from the abstract being processed is matched, are obtained by sorting the terms of the Authority List on first three characters. Within the subsets, the words are sorted by the number of significant characters in increasing order and alphabetically within the sub-subsets of terms with the same number of significant characters.

The subsets can be formed also by sorting the terms of the Authority List only on the first two characters instead of the first three, if the file

is not too large. Three characters are the upper limit for this purpose, since this is the lowest number of significant characters a term might be designed to have (there are no terms on the Authority List with two or one significant characters).

There is also a rule relating the number of significant characters against which a word is matched and the total number of characters in that word. This rule says that if the word is five characters long or less, it is matched only against those terms on the Authority List which have five significant characters or less. If the word is six characters long or longer, it is matched only against such terms of the Authority List which have five significant characters or more. Thus, the word

BATTERIES

in the text of an abstract would be matched against the Authority List term

BATTER*Y

on six significant characters and indexed by this term, but it would not be matched against the Authority List term

BAT*

on three significant characters, even if the Authority List would not contain BATTER*Y. Similarly, this rule would prevent the word DISCONTINUITY being accepted by the Authority list term DISC*, PUMPERNICKEL by PUMP, etc. This rule had to be applied because with the decreasing number of significant characters, the discriminating power of the Authority List terms with regard to longer words decreases very significantly.

The above described mechanism of the selection of significant terms performs at the same time the important function of editing the index by combining such similar terms which agree on the significant number of characters. Thus, as a result of the editing procedure, the abstracts containing the words

DIFFRACTION

DIFFRACTIONS

DIFFRACTED

DIFFRACTIVE

DIFFRACTS

would be posted under the index term DIFFRACTION and the abstracts containing the words

INHOMOGENEITY

INHOMOGENEITIES

INHOMOGENEOUS

INHOMOGENEOUSLY

would be posted under the index term INHOMOGENEITY. To give one more example abstracts containing

MAGNETIC

MAGNETICALLY

MAGNETIZE

MAGNETIZATION

MAGNET

MAGNETS

MAGNETO

MAGNETISM

would be posted under the index term MAGNETISM.

The programmed mechanism to provide human control and systems optimization capability (Residue Editing Program) consists of three subroutines:

a. Subroutine for the generation of the Residue Record with the frequency count of terms.

b. Subroutine for updating Kill List.

c. Subroutine for updating Authority List.

The subroutine for the generation of Residue Records produces a listing of words which do not match either with the Kill or with the Authority List. Furthermore, it counts the frequency of occurrence of such words and lists them in the decreasing order of occurrence. Basically, there could be three categories of words appearing on the Residue Record: words which are not acceptable as indexing terms, but which were not included in the Kill List, words which should have generated indexing terms but did not do so because there were no matching terms in the Authority List, and words which did not match with an existing term on the Kill or Authority List because of spelling errors.

The Residue Record is periodically reviewed by a human editor. In addition to correcting misspelled words, the human editor updates the index by adding the indexing terms derived from the significant terms and optimizes the system using the feedback for updating the Kill List and the Authority List.

In both cases, the frequency of appearance of the candidate terms for the Kill and Authority List in the Residue record serves as a criterion for creating new kill and authority terms. Thus, if an insignificant word appears reasonably often in the texts, it would be placed on the Kill List, but if such a word appears seldom, it might not be economically justifiable to create a new term for the Kill List, since this increases processing time. Similar considerations apply to the updating of the Authority List.

As a final product, the system delivers:

a. Subject Index to the documents in store. This is the index file sorted by indexing terms and, when printed, it gives indexing terms in alphabetical sequence with the accession numbers of documents to which these terms were assigned. The subject index can be produced with or without cross-references, depending on users requirements (See Annex X for a sample page).

b. Sets of indexing terms assigned to single documents. This is the index file sorted by accession numbers. The sets of indexing terms, or, as they are often referred to in this contract, the sets of key words would usually be printed with the abstracts, if such a print-out is at all required. (See Annex XI for a sample.)

Figures 7 through 11 show the flow charts of the system and its components.

FORMAL AUTO INDEXING OF SCIENTIFIC TEXTS (FAST) SYSTEM

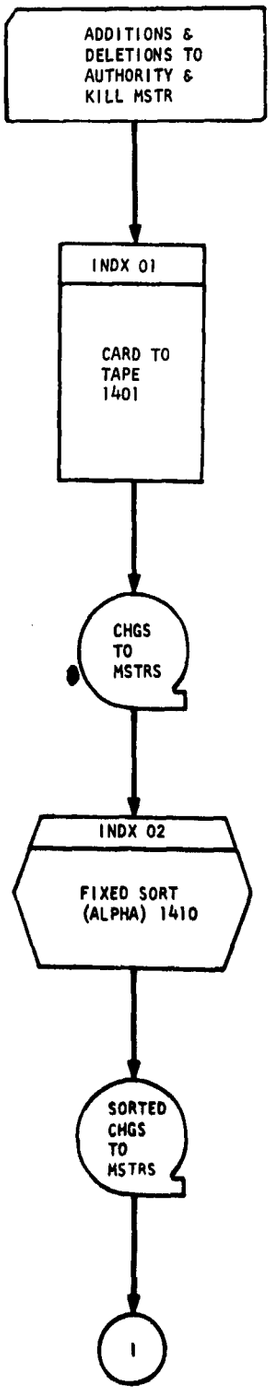

Figure 7

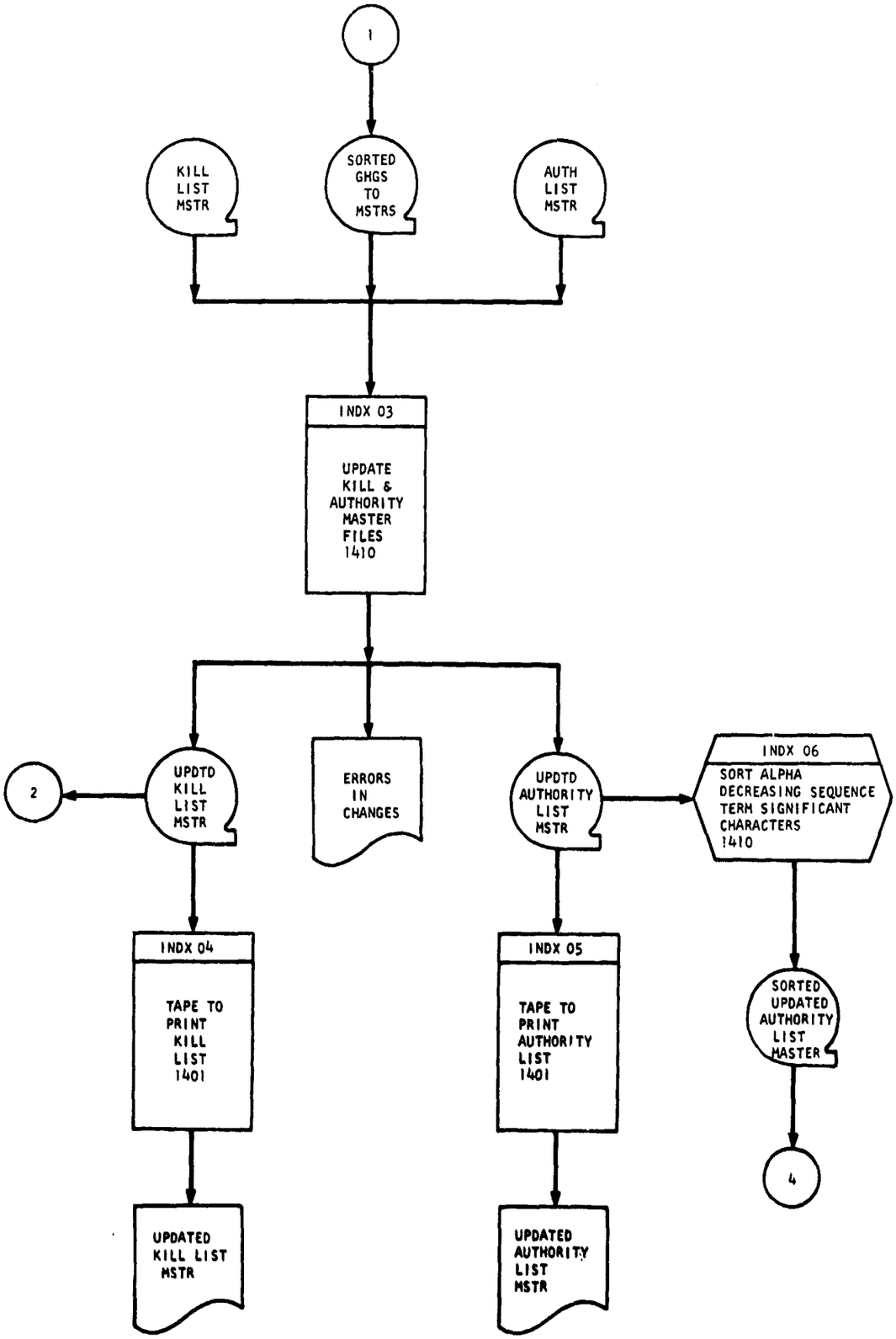
1
KILL LIST MSTR
SORTED GHGS TO MSTRS
AUTH LIST MSTR
INDX 03
UPDATE KILL & AUTHORITY MASTER FILES 1410
2
UPDTD KILL LIST MSTR
ERRORS IN CHANGES
UPDTD AUTHORITY LIST MSTR
INDX 06
SORT ALPHA DECREASING SEQUENCE TERM SIGNIFICANT CHARACTERS 1410
INDX 04
TAPE TO PRINT KILL LIST 1401
INDX 05
TAPE TO PRINT AUTHORITY LIST 1401
SORTED UPDATED AUTHORITY LIST MASTER
4
UPDATED KILL LIST MSTR
UPDATED AUTHORITY LIST MSTR

Figure 8

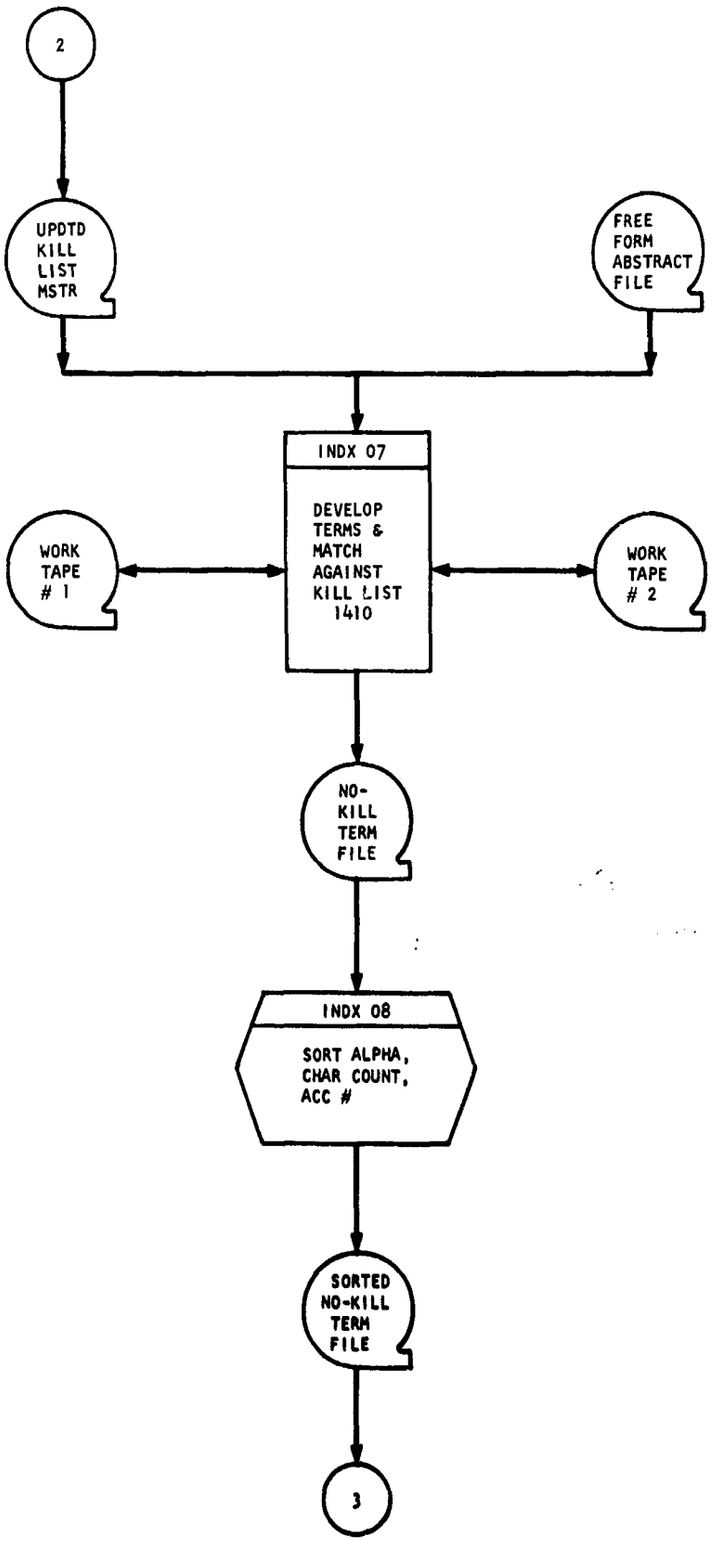
2
UPDTD KILL LIST MSTR
FREE FORM ABSTRACT FILE
INDX 07
DEVELOP TERMS & MATCH AGAINST KILL LIST 1410
WORK TAPE # 1
WORK TAPE # 2
NO-KILL TERM FILE
INDX 08
SORT ALPHA, CHAR COUNT, ACC #
SORTED NO-KILL TERM FILE
3

Figure 9

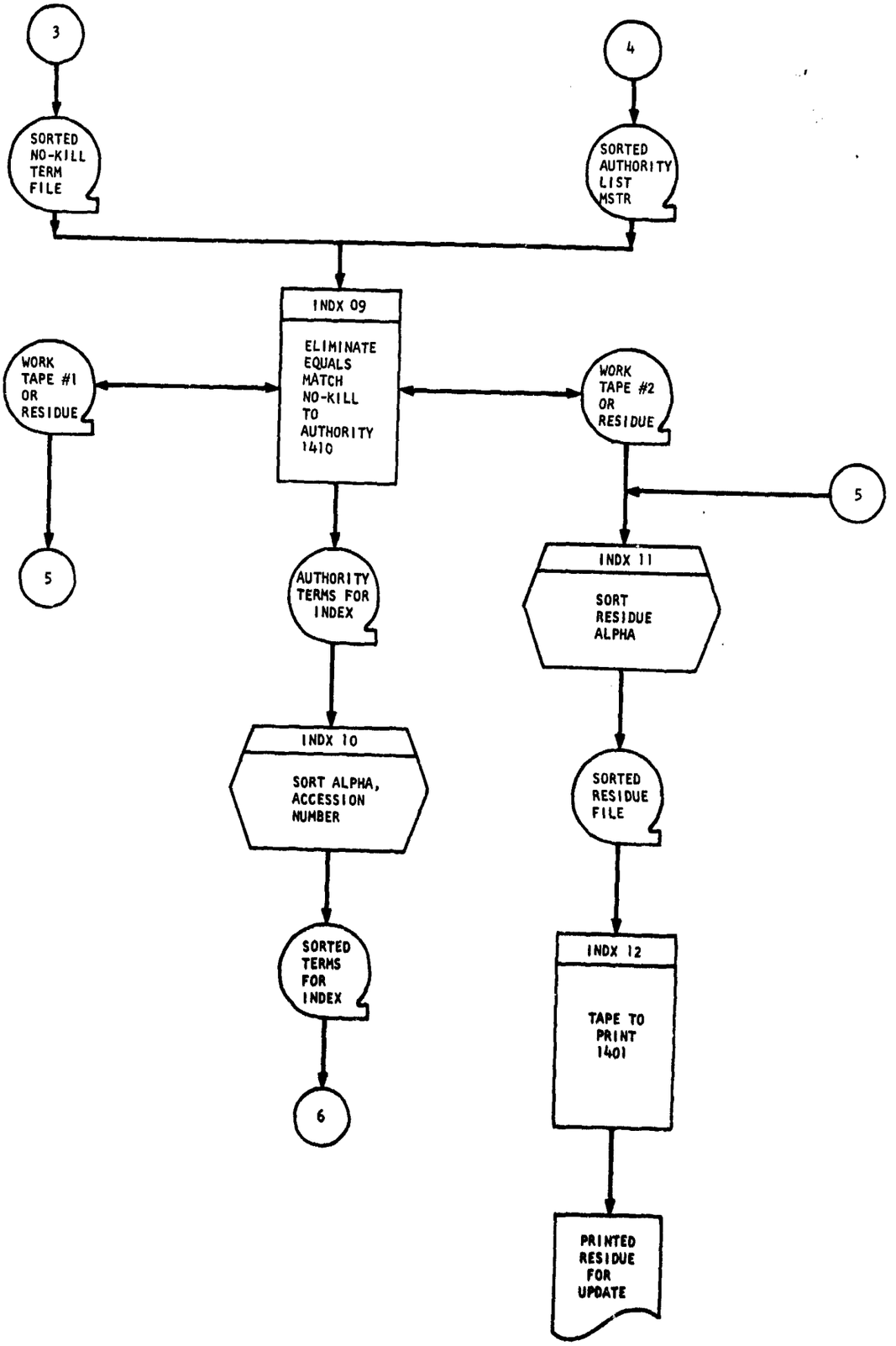
3
4
SORTED NO-KILL TERM FILE
SORTED AUTHORITY LIST MSTR
INDX 09
ELIMINATE EQUALS MATCH NO-KILL TO AUTHORITY 1410
WORK TAPE #1 OR RESIDUE
WORK TAPE #2 OR RESIDUE
5
5
AUTHORITY TERMS FOR INDEX
INDX 11
SORT RESIDUE ALPHA
INDX 10
SORT ALPHA, ACCESSION NUMBER
SORTED RESIDUE FILE
SORTED TERMS FOR INDEX
INDX 12
TAPE TO PRINT 1401
6
PRINTED RESIDUE FOR UPDATE

Figure 10

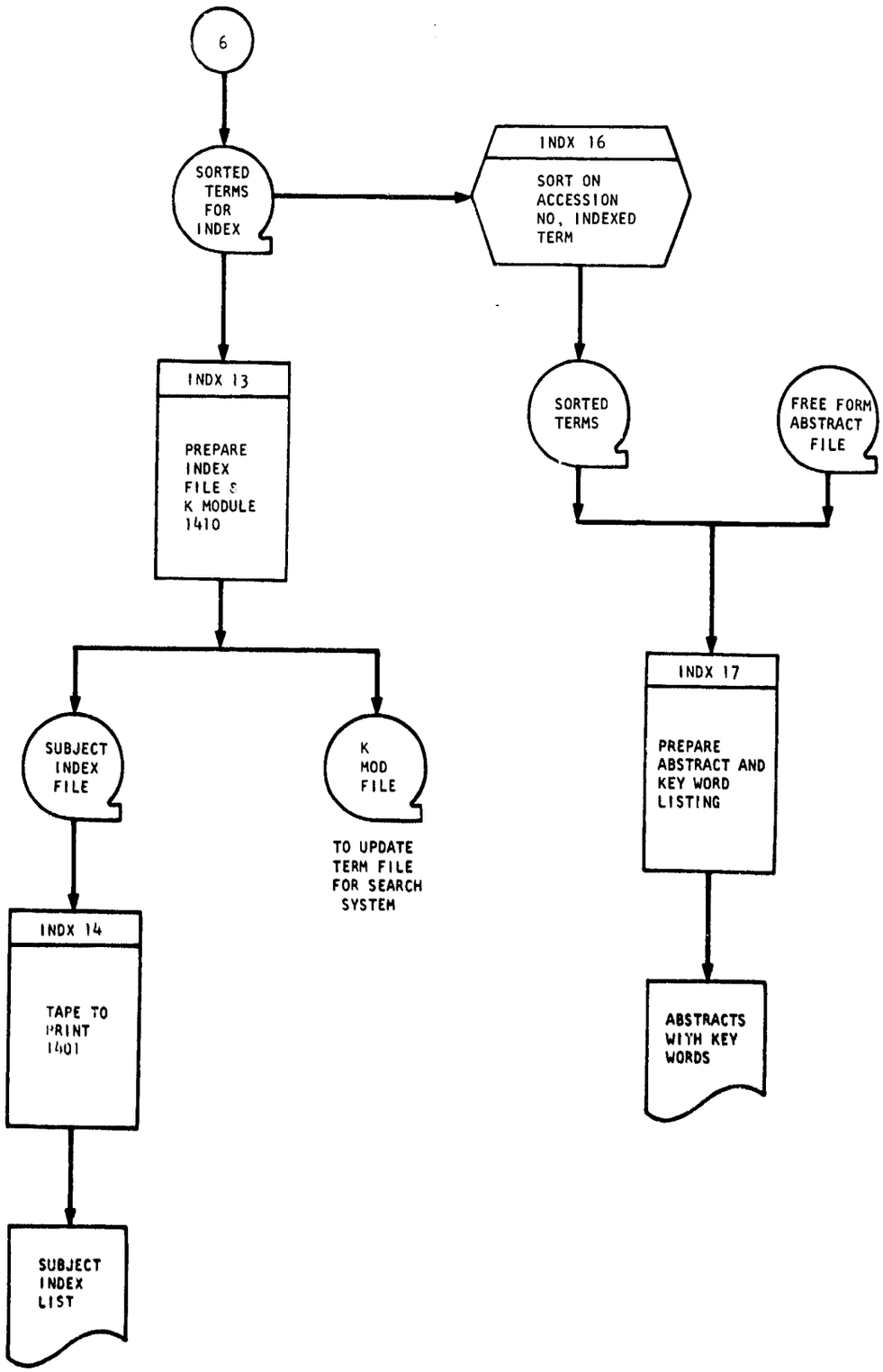
6
SORTED TERMS FOR INDEX
INDX 16
SORT ON ACCESSION NO, INDEXED TERM
INDX 13
PREPARE INDEX FILE & K MODULE 1410
SORTED TERMS
FREE FORM ABSTRACT FILE
INDX 17
PREPARE ABSTRACT AND KEY WORD LISTING
SUBJECT INDEX FILE
K MOD FILE
TO UPDATE TERM FILE FOR SEARCH SYSTEM
INDX 14
TAPE TO PRINT 1401
ABSTRACTS WITH KEY WORDS
SUBJECT INDEX LIST

Figure 11

II.4. CHARACTERISTICS OF THE INPUT INTO THE FAST SYSTEM

It was already mentioned that the FAST system was designed to process abstracts of scientific documents or short descriptions of research endeavors written in concise scientific language. This means that the system was optimized for this particular type of input. The length of a single item (abstract or task description) was to be approximately 200 words (see Samples in Annex I).

Three random samples were drawn from the total population of ILSE and OAR abstracts in store for a more detailed investigation of the characteristics of the input. The first sample contained 142 abstracts, the second and third contained 30 abstracts each. The average length of the documents used in actual tests of the system is given in Table 8.

Table 8: Number of documents and number of words in documents used in testing FAST system.

Sample No.	No. of Documents	Min. No. of Words in a Document	Max. No. of Words in a Document	Average No. Of Words Per Document
1	142	10	260	91.7
2	30	33	233	114.4
3	30	58	272	139.8

The total number of word occurrences in the documents of Sample No. 1 was 12,792. The number of different words in this population of word tokens*) was 2,841, so that on the average the same word occurred 4.503 times. Figures in Table 9 relate the number of different word types to the number of their occurrences in this document group.

A corresponding plot of the number of word types versus the number of their occurrences is shown in Figure 12 on a logarithmic scale.

The rank-frequency order of the 20 most frequent words in the above sample was as shown in Table 10.

E. S. Schwartz (231) reported 60 most frequent word types obtained after processing 10,000 and 19,710 word tokens from 7 popular magazine articles. The first 20 words from his list are reproduced in Table 11.

It is noted that only the rank 4 of the list in Table 10 and of the 10,000 token list of Table 11 identical as well as ranks 4 and 14 for the 19,710 token list. The ranks 1 through 10 of the words of the Sample No. 1 ILSE documents appear as ranks 2-1-5-4-6-3-30-(WILL is not included in the first

---

*) The term "word tokens" is used here in the sense of each word occurence in the text, some of which are exactly alike in their character structure, whereas "word types" is the subset of word tokens each one identifiable by a different character structure.

Table 9. Number of occurrences of word types in the population of 12,792 text words (word tokens) of sample No. 1 ILSE documents.

No. of Word Types	No. of Occurrences	No. of Word Types	No. of Occurrences
1	808	2	33
1	756	1	31
1	573	1	30
1	394	3	29
1	306	1	28
1	193	2	27
1	183	6	26
1	182	5	25
1	174	3	24
1	128	2	22
1	117	2	21
1	100	7	20
1	88	2	19
1	69	1	18
1	67	8	17
1	65	8	16
1	63	12	15
1	61	8	14
1	60	10	13
1	59	11	12
1	53	17	11
1	52	23	10
1	48	30	9
1	46	30	8
1	45	42	7
1	44	79	6
1	43	83	5
1	42	160	4
1	41	257	3
1	38	525	2
2	37	1,467	1
1	35		
		Σ =	2,841

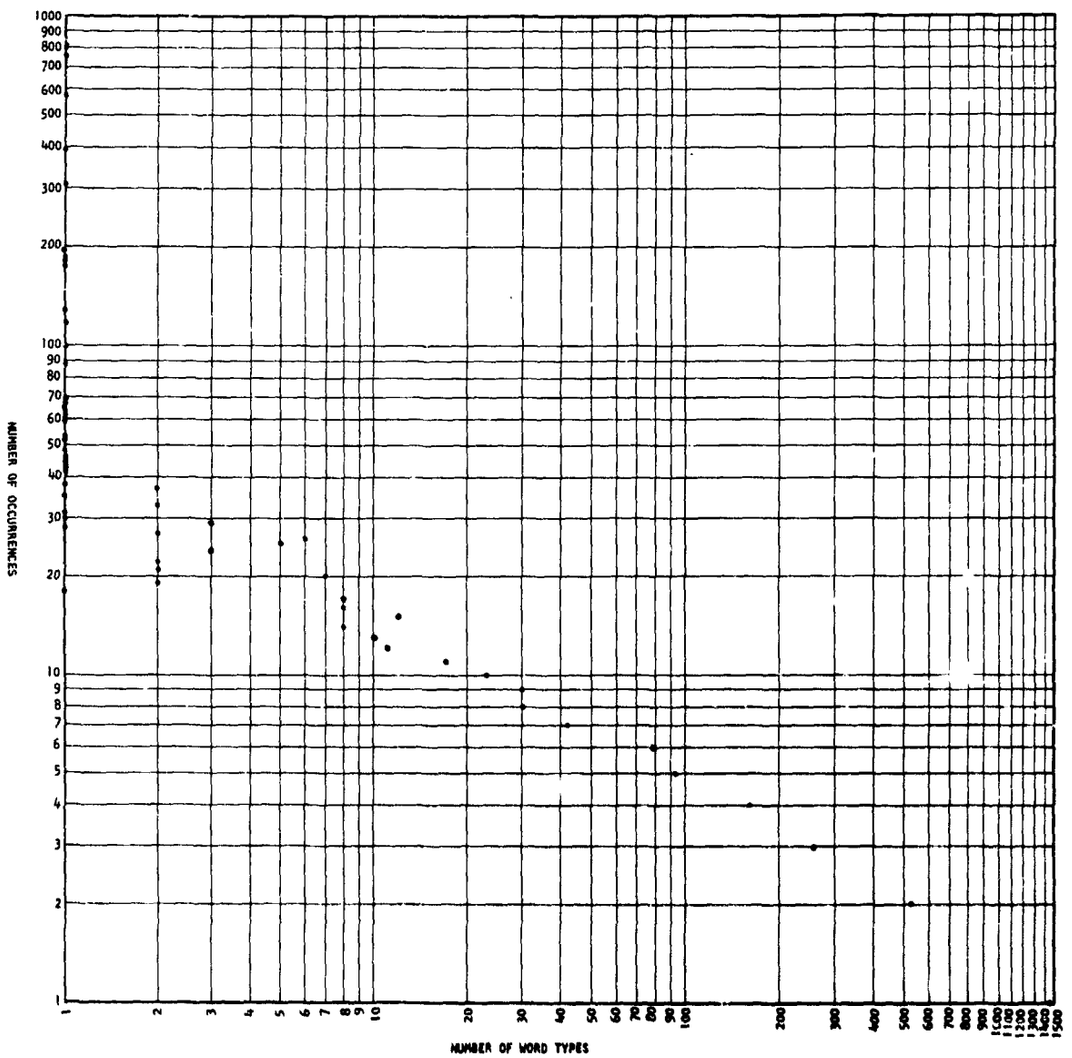

Figure 12: A plot of word frequencies versus word types for the sample No. 1 ILSE Documents.

Table 10. Rank-frequency order of the words in the sample No. 1 ILSE documents

Rank	Word Type	Frequency
1	OF	808
2	THE	756
3	AND	573
4	TO	394
5	IN	306
6	A	193
7	BE	183
8	WILL	182
9	FOR	174
10	THIS	128
11	IS	117
12	ON	100
13	AS	88
14	ARE	69
15	WITH	67
16	RESEARCH	65
17	STUDY	63
18	STUDIES	61
19	WHICH	60
20	BY	59

60 ranks)-11-26 on the 10,000 token list*) of Table 11 and as ranks 2-1-5-4-6-3-15-(WILL is again not included in the first 60 ranks)-11-27 on the 19,710 token list. The rank correlation between the top ten words (WILL in Table 10 is substituted by the next word) of the list is -4.106 and -1.292 respectively. The ten top words comprise 28.9 percent of word occurrences in the Sample No. 1 documents, whereas they comprise only 23 percent of the word occurrences in the 19,710 word sample. The 20 top words comprise 34.7 and 30 percent of word occurrences respectively.

*) Schwartz gives in his paper (231) as many as first 60 ranks of word types in order of their frequency.

Table 11. Rank-Frequency order of Word occurrences in 7 magazine articles.

10,000 Tokens			19,710 Tokens		
Rank	Word Type	Frequency	Rank	Word Type	Frequency
1	THE	657	1	THE	1192
2	OF	323	2	OF	677
3	A	274	3	A	541
4	TO	247	4	TO	518
5	AND	234	5	AND	462
6	IN	196	6	IN	450
7	THAT	109	7	THAT	242
8	IT	105	8	HE	105
9	HE	97	9	IS	190
10	IS	97	10	IT	181
11	FOR	79	11	FOR	157
12	WE	79	12	HIS	138
13	ON	75	13	ON	134
14	I	73	14	ARE	124
15	HIS	69	15	BE	123
16	WAS	64	16	WITH	121
17	THEY	62	17	I	112
18	YOU	62	18	HAVE	111
19	WITH	61	19	WAS	111
20	AS	59	20	YOU	106

Finally, the relation between the total number of word occurrences and the number of different words (word types) in ILSE documents was investigated and compared with available data on other texts. The data for this comparison were taken again from the above referenced paper of Schwartz (231). The results are summarized in Table 12.

Table 12. Word Counts by tokens and types

Study	Date	Material	Words		Percentage Of Types
			Tokens	Types	
Eldrige	1911	Newspaper articles	43,989	6,002	13.6
Dewey	1923	Miscellaneous	100,000	10,161	10.2
Hanley	1937	Joyce's "Ulysses"	260,430	29,899	11.5
Thorndike	1944	Miscellaneous	18,000,000	30,000	-
Miller-Newman	1958	Miscellaneous	36,299	5,537	15.2
Armour Research	1960	Military exercise	38,992	2,081	5.3
ILSE Documents	1965	Scientific task descriptions	12,792	2,841	22.2

Following conclusions can be derived from the above investigations:

1. By deleting duplicates, the population of words (word tokens) in ILSE type documents can be condensed to approximately 22 percent of its original volume.
2. The degree of the condensation thus achieved is less than for non-scientific texts or for texts not in abstract form.
3. The list of most frequent words in scientific abstracts and in articles from popular magazines differ considerably both in the rank order of identical word types and in the word types themselves.

II. 5. DESIGN AND TESTING OF SYSTEMS COMPONENTS

A. Kill List. The Kill List was designed to eliminate from the input data terms which:

(1) do not carry any information at all, such as words: of, the, but, are, have, etc.

(2) cannot be considered acceptable indexing terms because they possess little discriminatory power in the specific environment of their occurrence. For the particular type of documents processed, this category includes such terms as: RESEARCH, STUDY, TASK, etc.

On the other hand, a word might belong to one of the above described categories and yet not be placed on the Kill List because it does not appear often enough in the text to make such an inclusion desirable or economically justifiable. For one thing, certain limits as to the practical size of the list are set by the computer's memory capacity. Furthermore, checking whether a term on the Kill List appears in the text requires a certain amount of computer time, and if the possibility of such occurrences is low, it might be worth while to let it appear on the residue list of words which do not match either with the Kill List or with the Authority List. In other words, the final criterion for the inclusion of a term into the Kill List is a trade-off decision which takes into consideration the economics of computer processing time versus the economics of human editing of the no-match residue of the input data (Residue Record).

Samples No. 2 and No. 3, of 30 abstracts each, were processed against the Kill List containing 1,162 terms. This Kill List was derived from the Sample No. 1 of 142 abstracts. The results of the condensation of text thus achieved are shown in Table 13.

Table 13. Data on the processing of Samples No. 2 and No. 3 against the Kill List of 1,162 terms

	Number of Word Tokens Processed	No. of Word Tokens Eliminated By the Kill List	Percentage Of Reduction
Sample No. 2	3434	2,182	63.5
Sample No. 3	4194	2,332	55.6

The first seventy-three most frequent words eliminated from the word population of Sample No. 2 by processing against the Kill List are listed in Table 14.

The words in Table 14 account for 46.1 percent of all word occurrences in the documents of Sample 2. Thus the remaining 1,089 terms on the Kill List produced an additional reduction of the original volume of words of 17.4 percent only.

B. Authority List. It has been already mentioned that in addition to its prime function of selecting significant terms, the Authority File Program was designed also to combine conceptually related terms, which function corresponds to the human process of editing the index vocabulary. For conceptually related

Table 14. First seventy-three words deleted from the Sample No. 2 by processing against the Kill List in their order of frequency.

Word	No. of Occurrences	% of Total No. of Deletes
THE	242	11.0
OF	235	10.7
AND	130	5.9
TO	115	5.2
IN	107	4.9
A	46	2.1
IS	43	1.9
BE	36	1.6
THIS	35	1.6
ON	31	1.4
STUDY	29	1.3
WILL	29	1.3
STUDIED	26	1.1
BEEN	21	0.9
FOR	21	0.9
ARE	20	0.9
BY	19	0.8
RESEARCH	16	0.7
AS	15	0.6
DETERMINE	15	0.6
WHICH	14	0.6
FROM	13	0.5
OR	13	0.5
HAS	11	0.5
HAVE	11	0.5
AN	10	0.4
HUMAN	10	0.4
SYSTEMS	10	0.4
TASK	10	0.4
THAT	10	0.4
BEING	9	0.4
DURING	9	0.4
VARIOUS	9	0.4
AT	8	0.3
EFFECTS	8	0.3
MADE	8	0.3
PURPOSE	8	0.3
UNDER	8	0.3
HIGH	7	0.3
INVESTIGATION	7	0.3
OTHER	7	0.3

Table 14. (Continued)

Word	No. of Occurrences	% of Total No. of Deletes
TYPES	7	0.3
WAS	7	0.3
BETWEEN	6	0.2
CONDUCTED	6	0.2
FACTORS	6	0.2
NORMAL	6	0.2
SUCH	6	0.2
WOULD	6	0.2
CHANGES	5	0.2
DEVELOPMENT	5	0.2
EFFECT	5	0.2
INCLUDE	5	0.2
INTO	5	0.2
IT	5	0.2
MAN	5	0.2
SYSTEM	5	0.2
ASSOCIATED	4	0.1
BOTH	4	0.1
CERTAIN	4	0.1
FOUND	4	0.1
MEASURES	4	0.1
MORE	4	0.1
PROLONGED	4	0.1
PROVIDE	4	0.1
RELATIONSHIP	4	0.1
RELATIONSHIPS	4	0.1
THAN	4	0.1
THESE	4	0.1
USE	4	0.1
VARIABLES	4	0.1
WERE	4	0.1
YIELD	4	0.1

terms, which have certain characters in sequential order in common, this condensation and editing is achieved by matching on significant characters only. The reduction in the number of extracted significant words after their transformation into the new set of indexing terms (Uniterms) actually appearing in the subject index produced by FAST is shown in the Table 15.

Table 15. Reduction of the number of potential indexing terms for ILSE sample documents in the process of transformation by matching on significant characters.

	No. of Significant Words Before Transformation	No. of Indexing Terms After Transformation	% of Reduction
Sample No. 1	1,522	1,114	26.8
Sample No. 2	503	412	18.1
Sample No. 3	422	319	24.4

C. Residue Record. By regularly checking the Residue Record and updating both the Authority File and the Kill List as described in Section 6, it is possible to steadily reduce the number of words that were neither killed nor accepted by the Authority File Program. Specifically, by regular updating, it is possible to quickly reduce the number of significant words in the Residue Record, provided there are no essential changes in the subject field coverage of the documents processed. A sudden increase of significant words viz. potential indexing terms in the Residue Record unmistakable indicates that the input contains documents from a different field of knowledge than the one for which the system was primarily designed and optimized.

Table 16 gives numerical data on these residue records for the sample No. 2 and No. 3.

Table 16. Evaluation of the Residue Record (no-match output) for ILSE sample No. 2 and No. 3 documents

	Total No. of Word Occurrences on the Residue Record	Total No. of Word-Types on the Residue Record	No. of Word Types per Document	No. of Word Types Accepted as Indexing Terms per Document	No. of Word Types Rejected per Document
Sample					
No. 2	385	323	10.8	3.2	7.6
No. 3	726	551	18.4	0.6	17.8

11.6. DEPTH OF FAST INDEXING AND COMPARISON WITH HUMAN INDEXING

Machine generated indexing terms for each of the Sample No. 2 and No. 3 documents were compared with the indexing terms assigned to the same documents by human indexers for depth of indexing and commonality.

For the first set of 30 test documents, (sample No. 2) the FAST program assigned approximately twice as many indexing terms as the human indexers did. 57.3 percent of the indexing terms assigned by human indexers were picked also by the machine on the first run. After editing the residue and updating the Authority List, this figure increased to 65.2 percent. However, for that set of documents, these figures could not be considered unbiased because human indexers had information available which was not part of the input for automatic indexing process.

For the second set of 30 test documents (sample No. 3), the machine assigned approximately 46.4 percent more indexing terms per document than human indexers did. The respective figures o terms common with the terms selected by human indexers were 59.8 percent before update and 63.8 percent after update (see also Table 17).

The analysis of the terms, which were assigned by the human indexers but not by the FAST program, disclosed two major reasons for their appearance:

1. The indexers would assign more generic terms in addition to the terms in the abstract (e.g. HYDRODYNAMICS in addition to MAGNETOHYDRODYNAMICS when only the latter appeared in the text).
2. The indexers would assign synonymous terms (e.g. PLASMA when MAGNETOHYDRODYNAMICS appeared in text).

Table 17. Comparison of the depth of indexing and commonality for ILSE sample documents

Sample No.	No of Documents	Average No. of Words per Document	Average No. of Indexing Terms Assigned by FAST		Average No. of Indexing Terms Assigned by Humans	Percentage of Indexing Terms Assigned by Humans also Selected by FAST		Average No. of Unique FAST Selected Terms		Average No. of Unique Human Selected Terms		Average No. of Terms Common To FAST and Human	
			Before Update	After Update		Before Update	After Update	Before Update	After Update	Before Update	After Update	Before Update	After Update
Sample #2	30	114.4	20.6	23.2	11.7	57.3	65.2	13.9	15.6	5.0	4.1	6.7	7.6
Sample #3	30	139.8	20.8	21.5	14.2	59.8	63.8	12.3	12.5	5.7	5.2	8.5	9.0

In most cases this can be done also mechanically by more elaborate posting instructions in the computer program, if such a capability is required. Although the techniques for this were developed, they were not incorporated into the FAST program.

The relation between the total number of words in an abstract and the number of indexing terms assigned by the FAST program was also investigated. Annex IV, Items 1 and 2 give word counts by document with corresponding numbers of indexing terms assigned by the FAST program and the ratios of the number of indexing terms to the number of words in the documents. It was established that there is a strong rank correlation between the number of words in the document and the number of indexing terms assigned to that document. For sample No. 2, the rank correlation coefficient is 0.9106 and for sample No. 3, it is 0.[illegible]37 (See Table 18 and 19). However, the relation is not linear. This is clearly demonstrated by calculating the ratio of the indexing terms to the number of words in the document. Those ratios are plotted in the chart Figure 13. The rank correlation coefficients for these sets of values also indicate that there is practically no correlation between the document length in terms of number of words and the ratio of indexing terms to the number of words in a document. The rank correlation coefficients are 0.1 and -0.412 respectively (see Table 20 and 21).

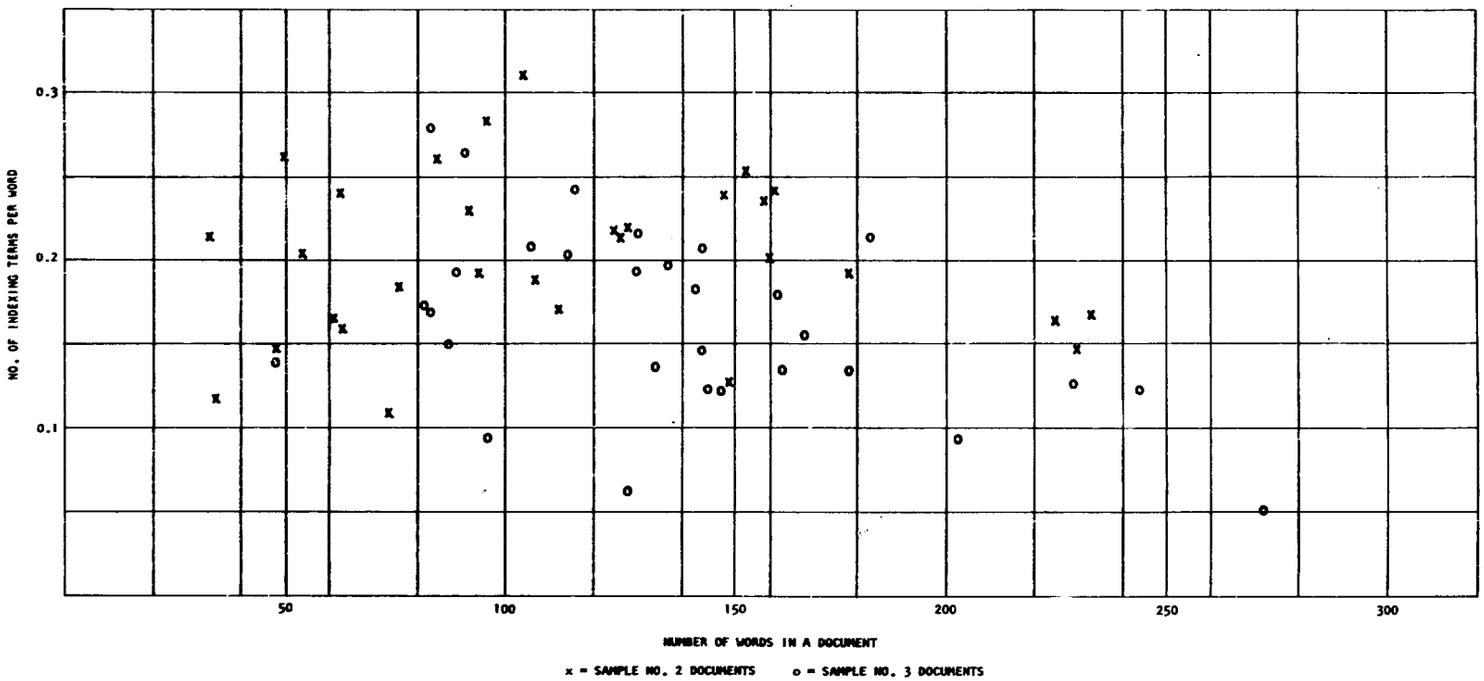

Figure 13: Plot of the number of indexing terms per word token of the document indexed versus total number of word tokens in the document.

Table 18. Rank correlation of document length (number of words in the document) to the number of indexing terms assigned by FAST for sample No. 2 documents (compare also Annex IV).

Document Length Rank (X_i)	Index Length Rank (Y_i)	$d_i = X_i - Y_i$	d_i^2
1	2	1	1
2	1	1	1
3	3	0	0
4	8	4	16
5	7	2	4
6	6	0	0
7	5	2	4
8	10	2	4
9	4	5	25
10	9	1	1
11	16	5	25
12	15	3	9
13	11	2	4
14	19	5	25
15	22	7	49
16	14	2	4
17	12	5	25
18	17	1	1
19	18	1	1
20	20	0	0
21	13	8	64
22	29	7	49
23	27	4	16
24	26	2	4
25	21	4	16
26	25	1	1
27	24	3	9
28	30	2	4
29	23	6	36
30	28	2	4
			$\sum d_i^2 = 40$

$$r' = 1 - 6\,\frac{\sum d_i^2}{N(N^2-1)} = 1 - \frac{6 \times 402}{30 \times 899} = 1 - \frac{2,412}{26,970} = 1 - 0.0894 = 0.9106$$

Table 19. Rank correlation of document length (number of words in the document) and the number of indexing terms assigned by the FAST for sample No. 3 documents (compare also Annex IV).

Document Length Rank (X_i)	Index Length Rank (Y_i)	$d_i = X_i - Y_i$	d_i^2
1	2	1	1
2	6	4	16
3	5	2	4
4	17	13	169
5	4	1	1
6	8	2	4
7	18	11	121
8	3	5	25
9	16	7	49
10	24	14	196
11	1	10	100
12	20	8	64
13	25	12	144
14	10	14	196
15	23	8	64
16	15	1	1
17	21	4	16
18	28	10	100
19	13	6	36
20	11	9	81
21	9	12	144
22	26	4	16
23	14	9	81
24	22	2	4
25	19	6	36
26	30	4	16
27	12	15	225
28	27	1	1
29	29	0	0
30	7	23	529
			$\Sigma = 2,440$

$$r' = 1 - \frac{6 \times 2,440}{26,970} = 1 - \frac{14,640}{26,970} = 1 - 0.543 = 0.457$$

Table 20. Rank correlation of document length (number of words in the document) and the ratio of the number of indexing terms to the number of words per document for sample No. 2 documents (Compare also Annex IV).

Document Length Rank (X_i)	Index Length Rank (Z_i)	$d = X_i - Z_i$	d_i^2
1	16	15	225
2	2	0	0
3	4	1	1
4	28	24	576
5	15	10	100
6	7	1	1
7	6	1	1
8	25	17	289
9	1	8	64
10	10	0	0
11	27	16	256
12	22	10	100
13	13	0	0
14	29	15	225
15	30	15	225
16	11	5	25
17	9	8	64
18	19	1	1
19	18	1	1
20	20	0	0
21	3	18	324
22	26	4	16
23	24	1	1
24	23	1	1
25	14	11	121
26	21	5	25
27	12	15	225
28	17	11	121
29	5	24	576
30	8	22	484
			$\sum d_i^2 = 4,048$

$$r' = 1 - 6\frac{\sum d_i^2}{N(N^2-1)} = 1 - \frac{6 \times 4,048}{30 \times 899} = 1 - \frac{24,288}{26,970} = 1 - 0.9 = 0.1$$

Table 21. Rank correlation of document length (number of words in the documents) and the ratio of the number of indexing terms to the number of words per document for sample No. 3 documents (Compare also Annex IV).

Document Length Rank (X_i)	Index Length Rank (Z_i)	$d_i = X_i - Z_i$	d_i^2
1	12	11	121
2	17	15	225
3	16	13	169
4	30	26	676
5	14	9	81
6	20	14	196
7	29	22	484
8	3	5	25
9	23	14	196
10	28	18	324
11	2	9	81
12	21	9	81
13	27	14	196
14	11	3	9
15	22	7	63
16	25	9	81
17	19	2	4
18	24	6	36
19	13	6	36
29	7	13	169
21	5	16	256
22	18	4	16
23	10	13	169
24	15	9	81
25	9	16	256
26	26	0	0
27	4	23	539
28	8	20	400
29	6	23	539
30	1	29	841
			$\Sigma = 6,350$

$$r' = 1 - \frac{6 \times 6,350}{26,970} = 1 - \frac{38,100}{26,970} = 1 - 1.412 = -0.412$$

II.7. INDEXING CONSISTENCY TESTS

Two types of consistency tests were made: inter-indexer and intra-indexer consistency tests. In this context, machine and author are considered "indexers".

The purpose of the inter-indexer consistency tests was to investigate the variation in the choice of indexing terms between two or more indexers (including author and machine) taken at a time. Six experienced indexers were given the same four documents to index. There was no communication among the indexers. They were not permitted to discuss the documents they indexed or to compare the terms they assigned. The documents were also indexed independently by the authors of the documents, and automatically by the FAST method. No effort was made to evaluate how good or bad were single indexing terms selected by the indexers, author or machine, since, in the investigator's opinion, there are no absolute and generally acceptable criteria for such an evaluation (assuming that the indexers possess the necessary amount of competence in their field). Consequently, the comparison was made on purely formal grounds.

The inter-indexer consistency coefficient was defined as the ratio of the number of terms which are common to a group of n individually recognizable indexers to the total number of different terms selected by these indexers. For a combination of n indexers*) at a time, the inter-indexer consistency coefficient is:

---

*) We remind again that the author and machine are also "indexers" for the purpose of this study.

$$\xi = \frac{T(A_1A_2 \ldots A_n)}{\sum_i T(A_i) - \sum'_{ij} T(A_iA_j) + \sum''_{ijk} T(A_iA_jA_k) - \ldots + (-1)\ T(A_1A_2 \ldots A_n)}$$

where

$T(A_1A_2 \ldots A_n)$ is the number of terms used by the indexers $A_1, A_2, \ldots A_n$ in common;

$T(A_i)$ - number of terms assigned to the document by the indexer A_i, $i = 1,2,\ldots n$;

$T(A_iA_j)$ - number of terms used by two indexers A_i and A_j $(i, j = 1,2,\ldots n, i \neq j)$ in common;

$T(A_iA_jA_k)$ - number of terms used by three indexers A_i, A_j and A_k in common etc.

and

$\sum'_{ij}$ means the sum over all i and j, with $i \neq j$

$\sum''_{ijk}$ means the sum over all i, j, k with no two of them equal, and

so on.

Obviously, if the indexers would all assign the same terms to a given document, then

$$T(A_1A_2 \ldots A_n) = \sum_i T(A_i) - \sum'_{ij} T(A_iA_j) + \ldots + (-1)^{n-1} T(A_1A_2 \ldots A_n)$$

and

$$\xi = 1$$

On the other hand, if the indexers would produce such sets of terms, that no elements (terms) were common for these sets, then

$$T(A_1A_2 \ldots A_n) = 0$$

and consequently

$$\xi = 0$$

As already mentioned, the indexers worked independently of one another and no consultation was permitted. Aside from the requirement that the documents should be indexed by the Uniterm method in order to be comparable with the machine generated indexes, there were no other restrictions imposed on the indexers: the indexers were not bound to a pre-established vocabulary, neither were they limited in the amount of indexing terms per document. Since the indexes created by the authors of the documents were not strictly Uniterm indexes, they were converted to Uniterm by breaking up pre-coordinated terms. The four documents which were thus indexed and the indexes evaluated for inter-indexer consistency, are reproduced in Annex V, Items 1 - 4.

Tables in Items 1 through 4, Annex VI, list the terms selected by various indexers, authors and machine, and the table in Annex VII gives consistency coefficients for various combinations of indexers for the four sample documents.

Table 22 gives the average values of consistency coefficients for the same four sample documents for different sizes of indexer groups. These figures reveal two very significant facts: (1) the substitution of the machine (FAST) for an experienced indexer does not significantly affect the inter-indexer consistency; the inter-indexer consistency of a group, one element of which is machine, rapidly approaches the inter-indexer consistency of a group of all-human indexers with the increasing number of elements (indexers) in the group whose products are compared. Furthermore, the figures in Table 22 show that the inter-indexer consistency is in all the cases higher if one human indexer is substituted by the machine (FAST) then when he is substituted by

the author. This can be considered a satisfactory proof of the adequacy of FAST indexing in comparison with human indexing.

Comparison was also made of variances for the set of data pertaining to combinations of two indexers or to indexer and machine, or indexer and author. These variances are given in Table 23. The average values of variances for all four sample documents are:

Two indexers	100.52×10^{-4}
Indexer - Author	226.52×10^{-4}
Indexer - Machine	28.26×10^{-4}

The above figures indicate, within certain confidence limits, the important fact, that the deviations from the mean consistency values are smaller when the sets of indexing terms produced by a human indexer are compared with corresponding sets produced by the FAST program than they are when sets of indexing terms produced by one human indexer are compared with those produced by another. In turn, the deviations from the mean consistency values are smaller for two human indexers than indexer and author comparisons. In other words, there are less drastic differences in selecting indexing terms for given documents between an indexer and the FAST program than between any two experienced indexers or between indexer and author.

In many practical cases the intra-indexer consistency is, however, even more important than the inter-indexer consistency. For the purpose of this study, the intra-indexer consistency is defined as the amount of consistency and reliability in selecting indexing terms when the same indexer re-indexes the same document after certain period of time. The time period

Table 22. Average values of consistency coefficients for indexer group sizes two through six for four sample documents.

Any 2 indexers	0.453	
One indexer & machine		0.392
One indexer & author		0.350
Any 3 indexers	0.307	
Any 2 indexers & machine		0.265
Any 2 indexers & author		0.213
Any 4 indexers	0.232	
Any 3 indexers & machine		0.207
Any 3 indexers & author		0.163
Any 5 indexers	0.187	
Any 4 indexers & machine		0.170
Any 4 indexers & author		0.133
Any 6 indexers	0.158	
Any 5 indexers & machine		0.144
Any 5 indexers & author		0.114

chosen was two months in order to reduce to a great extent the memory effects. The same four sample documents were used for the intra-indexer consistency test. Since, however, two of the indexers, who indexed the documents the first time, were no longer available for the re-indexing, only the results of four indexers were compared and evaluated. The test conditions for the re-indexing were the same as for the original indexing. Tables in Annex VII Items 1 through 4, show the indexing terms selected by the four indexers and by the FAST program during the first indexing round and during the re-indexing round. The indexing terms selected during the first round are checked "x" and those selected when the documents were re-indexed are checked by "0". Terms, which were picked both times, are checked by '⊗'.

Table 23. Variances $\hat{\sigma}^2$ of inter-indexer consistency coefficients for four sample documents.

	Document TSR No. 1	Document TSR No. 2	Document TSR No. 3	Document TSR No. 5
Two indexers	51.65×10^{-4}	126.50×10^{-4}	183.03×10^{-4}	40.92×10^{-4}
Indexer-Author	678.01×10^{-4}	55.25×10^{-4}	112.45×10^{-4}	60.38×10^{-4}
Indexer-Machine	55.99×10^{-4}	39.29×10^{-4}	13.81×10^{-4}	32.24×10^{-4}

The intra-indexing coefficient is defined as the ratio of the number of identical terms selected by the same indexer both first and second time to the total number of different terms used by the indexer.

Thus

$$\eta_c = \frac{T_{or}}{T_o + T_r - T_{or}}$$

where

T_{or} = number of same terms which have been used by the indexer both when indexing a document first time and re-indexing the same document after a lapse of time.

T_o = number of terms assigned by the indexer when the document was indexed first time.

T_r = number of terms assigned by the indexer when the document was re-indexed.

Obviously, if in re-indexing the document, an indexer would not assign any of the terms which he had assigned to the document when indexing it

for t. e first time, T_{or} would be equal to zero and also the intra-indexer consistency coefficient η_c would be equal to zero. On the other extreme, if an indexer would use exactly the same terms when re-indexing the document as he did the first time, then $T_{or} = T_o + T_r - T_{or}$ and the coefficient would be equal to 1.

Table 24 gives the intra-indexer coefficients for four human indexers and for the machine (FAST) calculated for each of the four sample documents indexed.

Table 24. Intra-indexer consistency coefficients for four sample documents.

	Document TSR No. 1	Document TSR No. 2	Document TSR No. 3	Document TSR No. 5
Indexer No. 1	0.750	0.643	0.765	0.642
Indexer No. 4	0.591	0.706	0.652	0.571
Indexer No. 5	0.500	0.666	0.600	0.590
Indexer No. 6	0.687	0.750	0.529	0.933
Machine (FAST Program)	1.000	1.000	1.000	1.000

The average intra-indexer consistency for all indexers and all tests was 0.661. This means that there is very high probability that an indexer will assign different sets of indexing terms to one and the same document at different points in time viz. that his judgment as to which terms are most representative of the contents of the document is not invariable,

but varies with time. Consequently, this results in a certain amount of uncertainity on behalf of the user as to the criteria which the indexers apply in selecting the indexing terms. The FAST Program, of course, performs always with 100% consistency.

II.8. CHANNEL CAPACITY AND EFFICIENCY

The ILSE 1963 Subject Index and the indexes produced by the FAST method for samples 2 and 3 of documents were also analyzed for the frequency and the distribution of postings under the subject terms. The table in Annex IX gives the frequency distribution of terms according to the number of postings or entries associated with these terms for ILSE 1963 Subject Index.

Figure 14 is a plot of the number of terms against the frequency of postings for each term group on the logarithmic scale paper for the same ILSE index. It can be noted that the plot in its general trend is somewhat similar to the Zipf-Mandelbrot curve of a log-log plot of word frequency versus word rank.*) However, because of much greater spread of single points, the difference is significant enough to prevent conclusion that the frequency of words as a function of word rank is equivalent to the frequency of postings as a function of term rank.

Houston and Wall (150) published statistics on ten indexed collections and plotted cumulative distributions of postings in these collections. They found that all these distributions are nearly log-normal. The plot is reproduced below on Figure 15. For the purpose of comparison, the cumulative distribution of postings for the ILSE 1963 system is also entered on the plot. Obviously the latter distribution follows the same general pattern

---

*) Zipf, G. K., 1949, Human Behavior and the Principle of Least Effort, Addision-Wesley Co., Inc., Cambridge, Mass.

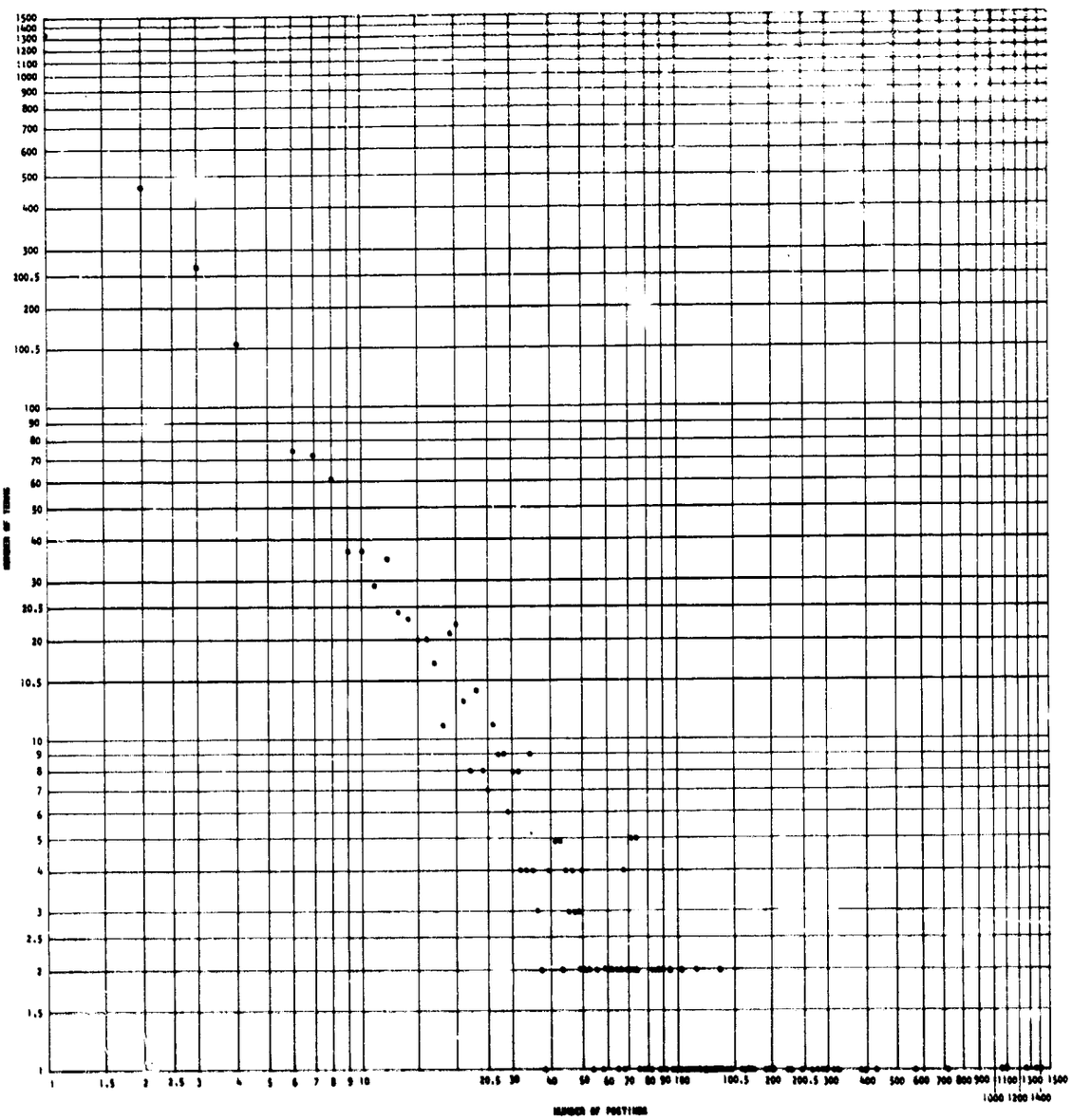

Figure 14: A plot of the number of terms versus frequency of postings on logarithmic scale for ILSE 1963 Subject Index.

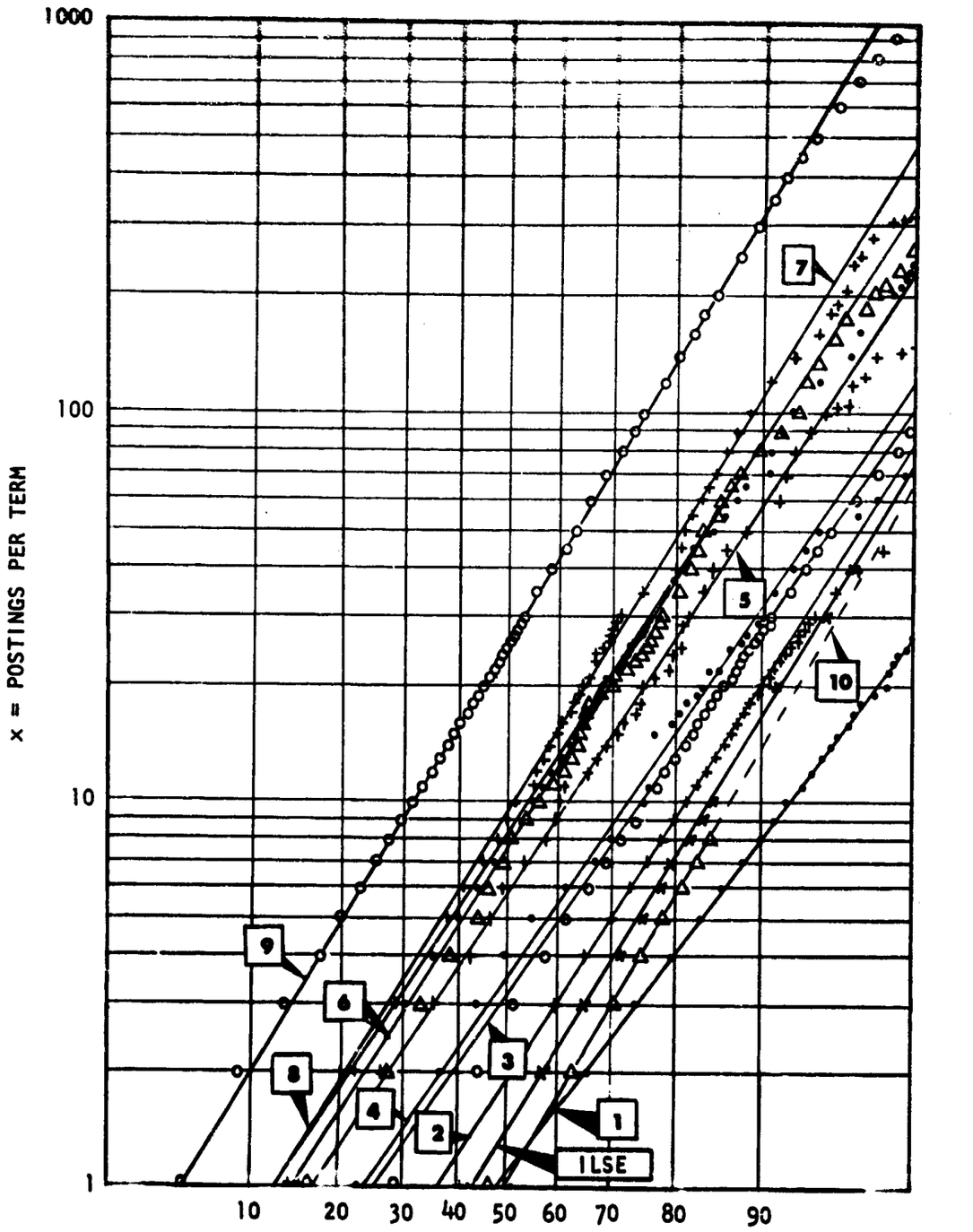

Figure 15: Cumulative distribution of postings per term for collections reported by Houston and Wall and for the ILSE 1963 Index.

as the ones reported by Houston and Wall, viz. could be considered as belonging to the class of log-normal distributions.

Obviously, indexes can be considered as channels for transmitting information from the store to the user. As such they can be formally evaluated by the information theory methods. This approach was suggested and investigated in greater detail by the author (281).

The overall efficiency of an index as an information channel can be expressed by the efficiency coefficient

$$\eta = \eta_I \cdot \eta_R$$

where η_I is the efficiency coefficient measuring the specificity of the information retrieved or the information content of the indexing terms and η_R is the efficiency coefficient measuring the retrievability and recall in terms of the operational economy of the system.
The coefficient η_I is obtained from the equation

$$\eta_I = \frac{\sum p_i \ln p_i}{C_I}$$

where

p_i is the probability of occurrence of the term P_i in the system or its relative frequency obtained as the ratio of the number of postings under this term to the total number of postings in the collection and C_I - channel capacity if the criterion is index specificity.

In this case

$$C_I = - \ln \frac{1}{P}$$

where P is the total number of postings in the index.

In a similar way, the coefficient

$$\eta_R = \frac{\sum g_i \ln g_i}{C_R}$$

where

g_i is the probability of a term having i number of postings ($i = 1,2,3....n$) or its relative frequency obtained as a ratio of the number of terms with i number of postings to the total number of terms

Table 25. Distribution of postings in indexes for sample No. 2 and No. 3 documents and collections.

Sample No. 2		Sample No. 3	
Number of Postings	Number of Indexing Terms	Number of Postings	Number of Indexing Terms
1	299	1	160
2	69	2	67
3	19	3	27
4	10	4	11
5	7	5	11
6	4	6	10
7	3	7	2
12	1	8	1
		9	4
		[illegible]	2
		12	1
		13	1
		14	1
	$\sum$ = 412		$\sum$ = 298

and

C_R - channel capacity if the criterion is retrievability based on operational economy. In this case

$$C_R = \bar{g} \cdot \ln \bar{g} - (\bar{g}-1) \cdot \ln(\bar{g}-1)$$

where

$\bar{g}$ is the average number of postings per term.

From the data given in the Table 25 we obtain.

Sample No. 2

$$\sum p_i \ln p_i = - 5.8343$$

$$C_I = \ln \frac{1}{626} = - 6.4393$$

$$\eta_I = \frac{5.8343}{6.4393} = 0.906$$

Sample No. 3

$$\sum p_i \ln p_i = - 5.399$$

$$C_I = \ln \frac{1}{651} = - 6.4785$$

$$\eta_I = \frac{5.399}{6.4785} = 0.8334$$

Sample No. 2

$$\sum g_i \ln g_i = 0.9288$$

$$C_R = \bar{g} \ln \bar{g} - (\bar{g}-1) \ln (\bar{g}-1) =$$

$$= 1.5194 \ln 1.5194 - 0.5194 \ln 0.5194 =$$

$$= 0.9755$$

$$\eta_R = \frac{0.9288}{0.9755} = 0.9521$$

Sample No. 3

$$\sum g_i \ln g_i = 1.4444$$

$$C_R = 2.1845 \ln 2.1845 - 1.1845 \ln 1.1845 =$$

$$= 2.1845 (7.6889 - 6.9077) +$$

$$- 1.1845 (7.0766 - 6.9077) =$$

$$= 1.7065 - 0.2001 =$$

$$= 1.5065$$

$$\eta_R = \frac{1.4444}{1.5065} = 0.9588$$

Thus we finally obtain the overall efficiency coefficient for Sample No. 2

$$\eta = \eta_I \cdot \eta_R = 0.906 \times 0.9521 = 0.8626$$

and for Sample No. 3

$$\eta = \eta_I \cdot \eta_R = 0.8334 \times 0.9588 = 0.7991$$

For the ILSE 1963 Subject Index, produced by human indexers, the corresponding coefficients are

$$\eta_I = 0.6095$$

$$\eta_R = 0.7314$$

and thus the overall efficiency coefficient is

$$\eta = \eta_I \cdot \eta_R = 0.6095 \times 0.7314 = 0.4458$$

Although the samples of the indexes produced by the FAST method, which were here investigated, were small to justify far reaching conclusions, they nevertheless indicate that such FAST indexes compare in efficiency very favorably to indexes produced by human indexers and that there is good reason to believe that they need less optimizing than the ones produced by the humans.

ANNEXES

ANNEX 1, ITEM 1

AGENCY- NASA ARC 0820 8

ZZ

TASK - STUDY OF LONG-TERM EFFECTS OF LOW G-LOADING OF MAMMALS

PRIN INV- DURATION- / / - / /

TO STUDY THE EFFECTS OF LONG-TERM EXPOSURE TO AN ALTERED G ENVIRONMENT /BY CENTRIFUGATION/ OF VARIOUS MAMMALS INCLUDING MICE, RATS. PHYSIOLOGIC AND BIOCHEMICAL EFFECTS WILL BE MEASURED TO DELINEATE THOSE RESPONSES WHICH ARE G-RESPONSIVE. CONTROL DATA AS WELL AS TEST ANIMAL DATA WILL ULTIMATELY BE APPLIED TO SETTING UP SPECIFIC EXPERIMENTS FOR SUSTAINED ZERO G STUDIES. ADAPTIVE CHANGES IN THE HOMEOSTATIC PROCESSES WILL BE FOLLOWED IN SUPRA ONE G ADAPTED ANIMALS WHEN THEY ARE RETURNED TO NORMAL G ENVIRONMENT. INTRACELLULAR EFFECTS OF SUSTAINED G LOADING WILL BE STUDIED PARTICULARLY CHANGES IN FAT AND CARBOHYDRATE METABOLISM OF MITOCHONDRIA AND PROTEIN METABOLISM OF ISOLATED MICROSOMAL FRACTIONS ALTERATIONS IN BLOOD AND TISSUE ISOENZYMES WILL BE STUDIED. METABOLIC STUDIES BOTH AT THE WHOLE ANIMAL LEVEL AS WELL AS THE TISSUE AND CELLULAR LEVELS WILL BE FOLLOWED WITH LABELED SUBSTRATES. PROCESSES INVOLVED IN ADAPTING ANIMALS TO G-LOADS GREATER THAN ONE G WILL BE STUDIED AS WELL AS THE REVERSE PROCESS IN SUPRA ONE G ADAPTED ANIMALS.

ANNEX 1, ITEM 2

AGENCY- NASA ARC 0819 8

ZZ

TASK - NEUROHORMONAL STUDIES AS RELATED TO SPACE FLIGHT STRESSES

PRIN INV- DURATION- / / - / /

NEUROHORMONAL ASPECTS OF BRAIN MECHANISMS AND STRESS. /1/ TO IDENTIFY THE NEUROHORMONE FROM THE HYPOTHALAMUS WHICH RELEASES ACTH FROM THE PITUITARY. EVIDENCE SO FAR INDICATES THAT THIS IS VASOPRESSIN /ADH/. /2/ TO ASSAY VASOPRESSIN IN BRAIN TISSUE, IN JUGULAR BLOOD AND IN C-S FLUID IN ANIMALS UNDER VARIOUS PHYSIOLOGICAL AND UNPHYSIOLOGICAL CONDITIONS SUCH AS PHYSICAL AND PSYCHOLOGICAL STRESSES. /3/ TO INVESTIGATE THE MECHANISMS BY WHICH VASOPRESSIN IS RELEASED FROM THE HYPOTHALAMUS UNDER STRESS AND ROLE OF VASOPRESSIN IN THE SYNTHESIS AND DEGRADATION OF ACTH /WITH DR. STANLEY ELLIS/. /4/ TO MEASURE ADRENAL STEROIDS AND CATECHOLAMINES IN BLOOD & URINE IN ANIMALS AND MAN UNDER STRESS CONDITIONS.
SUBJECTING THE ORGANISM /INCLUDING MAN/ TO UNDUE STRESS SUCH AS THE PHYSICAL STRESS OF ACCELERATION, DECELERATION, WEIGHTLESSNESS, VIBRATION AND RADIATION AND TO PSYCHOLOGICAL AND PHYSIOLOGICAL STRESSES SUCH AS CONFINEMENT IN A SATELLITE, ANXIETY, DISTURBANCES IN SLEEP AND BIOLOGICAL RHYTHMS, FATIGUE, PAIN AND OTHER BODILY DISCOMFORTS, MAY SEVERELY CHALLENGE THE HOMEOSTATIC MECHANISMS OF THE BODY. IT IS IMPORTANT TO KNOW WHAT HAPPENS TO MAN IF THE HIGHER OR LOWER LIMITS OF THESE REGULATORY MECHANISMS ARE PASSED OVER. CALLING ATTENTION TO THESE FUNCTIONS SERVES TO EMPHASIZE THE IMPORTANCE OF STUDYING THE **TRIGGER** MECHANISM IN THE HYPOTHALAMUS WHICH GAVE OUT THE EARLIEST SIGNAL OF STRESS.

2	AMMONIA
2	AMMONIUM
1	AMPEROMETRY
1	AMPHETAMINE
2	AMPHIBIA
1	AMPHIPODA
1	AMPLIFICATION
8	AMPLIFIER
3	AMPLITUDE
2	AMPUTATED
1	ANABOLIC
6	ANAEROBIC
1	ANALGESIA
26	ANALOG
250	ANALYSIS
6	ANALYZER
2	ANAPHYLAXIS
15	ANATOMY
2	ANEMIA
1	ANEMONE
10	ANESTHESIA
1	ANEURYSM
3	ANGLE
8	ANGULAR
1	ANHYDRASE
317	ANIMAL
2	ANIMATION
1	ANION
1	ANNOYANCE
2	ANOXIA
1	ANSERINE
1	ANTARCTIC
1	ANTENNA
2	ANTERIOR
1	ANTHRANILIC
1	ANTHROPOID
6	ANTHROPOLOGY
26	ANTHROPOMETRY
1	ANTHROPOMORPHIC
10	ANTIBIOTICS
26	ANTIBODY
1	ANTICIPATORY
1	ANTIDIURETIC
7	ANTIDOTE
2	ANTIFEBRILE
1	ANTIFOG
18	ANTIGEN
1	ANTIHISTAMINE
1	ANTIHORMONE
1	ANTIOXIDANT
1	ANTIPITUITARY
1	ANTIPYRETIC

1	ANTIPYRINE
1	ANTIRADIATION
3	ANTISERUM
2	ANTISUBMARINE
1	ANTITHYROID
1	ANTITHYROTOXIC
1	ANTITUMOR
3	ANXIETY
1	AORTIC
1	APEX
2	APLYSIA
23	APOLLO
4	APPARATUS
1	APPETITE
2	APPLICATION
1	APPORTIONMENT
2	APPROACH
1	APPROXIMATING
6	APTITUDE
4	AQUATIC
1	ARABINOSUS
2	ARC
2	ARCTIC
2	AREA
1	ARGON
1	ARID
1	ARITHMETIC
1	ARMAMENTARIUM
3	ARMED
3	ARMY
1	AROMATIC
3	AROUSAL
1	ARTERIO-RENOUS
18	ARTERY
1	ARTHROPOD
12	ARTIFICIAL
1	ASCARIS
3	ASCORBIC
1	ASH
1	ASIA
1	ASPARAGINE
1	ASPECT
3	ASPERGILLUS
1	ASPHYXIATION
1	ASPIRIN
3	ASSAY
6	ASSEMBLY
1	ASSESSMENT
2	ASSIGNMENT
2	ASSIMILATION
1	ASSOCIATED
2	ASSOCIATION

137	MANNED
8	MANNING
1	MANNITOL
1	MANNURONIC
3	MANOMETRY
1	MANPOWER
8	MANUAL
1	MANUFACTURING
1	MANUSCRIPT
1	MAP
1	MAPPING
2	MARGINAL
74	MARINE
1	MARINELAND
1	MARK
1	MARKER
3	MARMOSET
12	MARROW
20	MARS
7	MASK
24	MASS
3	MASTER-SLAVE
1	MATCH-TO-SAMPLE
3	MATCHING
39	MATERIAL
56	MATHEMATICS
1	MATING
3	MATTER
1	MEAL
2	MEAN
2	MEANINGFUL
61	MEASUREMENT
8	MEAT
13	MECHANICAL
6	MECHANICS
59	MECHANISM
1	MECHANOCHEMISTRY
4	MECHANODYNAMICS
3	MEDIATION
288	MEDICINE
2	MEDITERRANEAN
9	MEDIUM
4	MEDULLA
1	MEETING
1	MEMBER
27	MEMBRANE
15	MEMORY
1	MENINGOENCEPHALITIS
1	MENINGOPNEUMONITIS
7	MENTAL
1	MENTALITY
1	MERCAPTAN

4	MERCURY
1	MESCALINE
1	MESOPHILIC
1	MESS
2	MESSAGE
165	METABOLISM
6	METAL
1	METAPLASIA
1	METASTABLE
6	METEORITE
3	METEOROID
1	METEOROLOGY
5	METER
1	METHIODIDE
2	METHIONINE
41	METHOD
4	METHODOLOGY
2	METHYL
1	MEXICO
5	MICROANALYSIS
2	MICROBALLOON
7	MICROBE
1	MICROBEAM
99	MICROBIOLOGY
2	MICROCALORIMETRY
1	MICROCHEMISTRY
1	MICROCHROMATOGRAPHY
3	MICROCONTAMINANT
1	MICRODOSIMETRY
1	MICROELECTRICITY
5	MICROELECTRODE
2	MICROELECTRONICS
1	MICROELECTROPHYSIOLOGY
2	MICROFLORA
1	MICROFLUOROMETER
1	MICROLEPIDOPTEREAN
46	MICROORGANISM
3	MICROPHONE
1	MICROPHYSIOLOGY
1	MICROPIPETTE
28	MICROSCOPE
16	MICROSCOPY
1	MICROSOME
2	MICROSPECTROPHOTOMETRY
1	MICROSPECTROSCOPY
1	MICROSTIMULUS
1	MICROSPECTROGRAPHY
3	MICROTECHNIQUE
17	MICROWAVE
1	MICRURGY
5	MIDDLE
1	MIDOCEAN

1	SOUNDING
11	SOURCE
1381	SPACE
18	SPACEPLANE
1	SPASM
6	SPATIAL
3	SPECIAL
2	SPECIALTY
1	SPECIES
5	SPECIFICITY
19	SPECIMEN
2	SPECTROFLUOROMETER
2	SPECTROGRAPHY
15	SPECTROMETRY
1	SPECTROPHOTOMETRIC
13	SPECTROPHOTOMETRY
15	SPECTROSCOPY
24	SPECTRUM
18	SPEECH
6	SPEED
1	SPHAEROTILUS
1	SPICARIA
2	SPIDER
1	SPILLAGE
4	SPIN
6	SPINAL
1	SPIROCHETE
4	SPIRODELLA
1	SPIROMETER
1	SPIROPYRAN
1	SPLEEN
1	SPLENIC
1	SPONGE
1	SPONTANEOUS
5	SPORE
3	SPORULATION
1	SPRAY
1	SQUIB
1	SQUID
11	STABILITY
1	STAGE
3	STAINING
6	STANDARD
4	STANDARDIZATION
1	STANDING
1	STAPHYLOCOCCUS
1	STAR
11	STATE
5	STATE-OF-THE-ART
1	STATIC
48	STATION
41	STATISTICS

7	STEM
22	STERILIZATION
9	STEROID
59	STIMULATION
1	STIMULATOR
1	STIMULI-FREE
2	STIMULUS
3	STOCHASTIC
1	STOP
47	STORAGE
2	STRAIN
1	STRAIN-GAUGE
1	STRAINING
4	STRATEGY
3	STRATOSPHERE
4	STREAM
4	STRENGTH
3	STREPTOCOCCUS
1	STREPTOLYSIN
1	STREPTOMYCES
267	STRESS
1	STRIKE
1	STRIP
1	STRONTIUM
45	STRUCTURE
1	STRUCTURE-IN-INTERACTION
4	STUDENT
22	STUDY
1	SUB-GROUPING
12	SUBCELLULAR
1	SUBCLASS
2	SUBCLINICAL
1	SUBCORTICAL
2	SUBGRAVITY
13	SUBJECT
1	SUBJECTIVE
7	SUBMARINE
1	SUBMICROSECOND
1	SUBSONIC
5	SUBSTITUTE
3	SUBSTRATE
17	SUBSYSTEM
1	SUBTILIS
4	SUCCESS
2	SUCCESSION
1	SUCCINIC
1	SUDAN
2	SUDOMOTOR
1	SUFFOCATION
8	SUGAR
70	SUIT
10	SULFATE

TERM AS PRINTED IN SUBJECT INDEX	SIGNIFICANT CHARACTERS	
≠		
*** CHARACTER CT 430 ***		
MANIPULATOR	11MANIPULATOR	≠
MANNED	06MANNED	≠
MANNITOL	06MANNIT	≠
MANNURONIC	06MANNURON	≠
MANOMETRY	06MANOMET	≠
MANPOWER	06MANPOW	≠
MANUFACTURING	06MANUFACTUR	≠
MANUSCRIPT	06MANUSCR	≠
MAP	03MAP	≠
MAPPING	06MAPPIN	≠
≠		
*** CHARACTER CT 430 ***		
MARAGING	06MARAGI	≠
MARBAC	06MARBAC	≠
MARINE	06MARINE	≠
MARINELAND	08MARINELAND	≠
MARK	04MARK	≠
MARKER	06MARKER	≠
MARMOSET	06MARMOSET	≠
MARROW	06MARROW	≠
MARS	04MARS	≠
MARTENSITE	06MARTENSIT	≠
≠		
*** CHARACTER CT 430 ***		
MARTIAN	06MARTIA	≠
MASER	05MASER	≠
MASK	04MASK	≠
MASKING	05MASKI	≠
MASS	04MASS	≠
MASTER-SLAVE	09MASTER-SLAV	≠
MATCH-TO-SAMPLE	13MATCH-TO-SAMP	≠
MATCHING	06MATCHI	≠
MATHEMATICS	06MATHEMAT	≠
MATING	06MATING	≠

≠
*** CHARACTER CT 430 ***

MATRIX	05MATRI	≠
MEAL	04MEAL	≠
MEASUREMENT	08MEASUREM	≠
MEAT	04MEAT	≠
MECHANICS	08MECHANIC	≠
MECHANISM	08MECHANIS	≠
MECHANOCHEMISTRY	09MECHANOCHEM	≠
MECHANODYNAMICS	09MECHANODYN	≠
MEDIATION	06MEDIAT	≠
MEDICINE	05MEDIC	≠

Word counts by document with corresponding numbers of indexing terms assigned by the FAST program and ratios of the number of indexing terms to the number of words in the documents of Sample No. 2. Figures in brackets show the ranks.

Document Log No.	Words Total	No. of Indexing Terms Assigned by FAST	No. of Indexing Terms per 10 Words
2820	34 (2)	4 (1)	1.18 (2)
2825	85 (11)	22 (16)	2.59 (27)
2861	125 (18)	27 (17)	2.16 (19)
2904	230 (29)	34 (23)	1.48 (5)
2908	233 (30)	39 (28)	1.67 (8)
3150	159 (24)	37 (26)	2.33 (23)
3141	63 (7)	10 (5)	1.59 (6)
3167	74 (9)	8 (4)	1.08 (1)
3203	107 (16)	20 (14)	1.87 (11)
3234	155 (22)	39 (29)	2.52 (26)
3342	92 (12)	21 (15)	2.28 (22)
3400	225 (28)	48 (30)	2.13 (17)
3413	178 (27)	34 (24)	1.91 (12)
3426	112 (17)	19 (12)	1.70 (9)
3779	126 (19)	27 (18)	2.14 (18)
3853	63 (8)	15 (10)	2.38 (25)
3856	161 (26)	36 (25)	2.24 (21)
4037	33 (1)	7 (2)	2.12 (16)
4276	157 (23)	37 (27)	2.36 (24)
4316	151 (21)	19 (13)	1.26 (3)
4846	96 (14)	27 (19)	2.81 (29)
5121	54 (5)	11 (7)	2.04 (15)
5315	50 (4)	13 (8)	2.60 (28)
5332	61 (6)	10 (6)	1.64 (7)
5351	48 (3)	7 (3)	1.46 (4)
5449	94 (13)	18 (11)	1.91 (13)
5481	160 (25)	32 (21)	2.00 (14)
5544	128 (20)	28 (20)	2.19 (20)
5833	76 (10)	14 (9)	1.84 (10)
5957	104 (15)	32 (22)	3.08 (30)

Word counts by document with corresponding numbers of indexing terms assigned by the FAST program and ratios of the number of indexing terms to the number of words in the documents of Sample No. 3. Figures in brackets show the ranks.

Document Log No.	Words Total	No. of Indexing Terms Assigned by FAST	No. of Indexing Terms per 10 Words
308375	91 (7)	24 (18)	2.64 (29)
308392	145 (18)	30 (28)	2.07 (24)
308455	178 (25)	24 (19)	1.35 (9)
308416	244 (29)	30 (29)	1.23 (6)
308384	130 (12)	25 (20)	1.92 (21)
207363	162 (22)	29 (26)	1.79 (18)
207362	143 (17)	26 (21)	1.82 (19)
207361	163 (23)	22 (14)	1.35 (10)
207371	183 (26)	39 (30)	2.13 (26)
207368	89 (6)	17 (8)	1.91 (20)
006691	128 (11)	8 (1)	0.62 (2)
006721	114 (9)	23 (16)	2.02 (23)
003175	83 (3)	14 (5)	1.69 (16)
004348	229 (28)	29 (27)	1.27 (8)
003494	149 (21)	18 (9)	1.21 (5)
308374	82 (2)	14 (6)	1.71 (17)
308397	168 (24)	26 (22)	1.55 (15)
308388	137 (15)	27 (23)	1.97 (22)
308421	106 (16)	22 (15)	2.07 (25)
308454	145 (19)	21 (13)	1.45 (13)
207313	96 (8)	9 (3)	0.94 (3)
207306	134 (14)	18 (10)	1.35 (11)
207386	58 (1)	8 (2)	1.38 (12)
207421	87 (5)	13 (4)	1.49 (14)
000619	83 (4)	23 (17)	2.77 (30)
004400	146 (20)	18 (11)	1.23 (7)
006788	272 (30)	14 (7)	0.51 (1)
001131	116 (10)	28 (24)	2.41 (28)
002408	203 (27)	19 (12)	0.94 (4)
006758	130 (13)	28 (25)	2.15 (27)

UNCLASSIFIED
Security Classification

ANNEX V ITEM 1

DOCUMENT CONTROL DATA - R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)
Plasma Laboratory (R. W. Gould)
California Institute of Technology
Pasadena, California

2a. REPORT SECURITY CLASSIFICATION

2b. GROUP

3. REPORT TITLE

An Experimental Study of Compressional Hydromagnetic Waves

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)
Technical Report No. 1, June 1963

5. AUTHOR(S) (Last name, first name, initial)

SWANSON, D. G.

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
June 1963	105	20

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
Grant No. 412-63	
b. PROJECT NO.	TR No. 1
c.	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
d.	

10. AVAILABILITY/LIMITATION NOTICES

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
	Office of Aerospace Research A. F. Office of Scientific Research

13. ABSTRACT. An experiment is described in which a compressional hydromagnetic wave is observed in a hydrogen plasma-filled waveguide. The theory of a cool, partially ionized, resistive plasma in a magnetic field is described briefly and expressions are derived for the dispersion relation and transfer function which include both the propagation and attenuation constants as a function of frequency. Measurements of the cutoff frequency are presented which verify its linear dependence on the magnetic field, and they show good agreement with theory on the variation with the ion mass density. The impulse response of the plasma is studied, transformed into the frequency domain, and quantitative comparisons are made with the theoretical transfer function to determine the degree of ionization, the resistivity, and the ion-neutral collision frequency.

Results indicate that the degree of ionization varies over a range from 75% to 45% when the initial density changes from $1.3 \cdot 10^{21}$ to $1.4 \cdot 10^{22}$ atoms/m^3. The measured resistivity appears to increase with the magnetic field, with the mean value corresponding to a temperature of the order of $5 \cdot 10^3$ °K. The average value of the product of the charge exchange cross section and the neutral thermal speed is found to be approximately $(5.5 \pm 1.3) \cdot 10^{-15}$ m^3/sec.

DD FORM 1 JAN 64 1473

Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
impulse response of a plasma						
resistivity measurement from damping of hydromagnetic waves						
plasma-filled waveguide						
plasma density measurement						
hydromagnetic wave propagation(magnetohydrodynamic waves)						
measurements in a decaying hydrogen plasma						
plasma wave propagation						

INSTRUCTIONS

1. ORIGINATING ACTIVITY: Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization *(corporate author)* issuing the report.

2a. REPORT SECURITY CLASSIFICATION: Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2b. GROUP: Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. REPORT TITLE: Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. DESCRIPTIVE NOTES: If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. AUTHOR(S): Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. REPORT DATE: Enter the date of the report as day, month, year, or month, year. If more than one date appears on the report, use date of publication.

7a. TOTAL NUMBER OF PAGES: The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7b. NUMBER OF REFERENCES: Enter the total number of references cited in the report.

8a. CONTRACT OR GRANT NUMBER: If appropriate, enter the applicable number of the contract or grant under which the report was written.

8b, 8c, & 8d. PROJECT NUMBER: Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9a. ORIGINATOR'S REPORT NUMBER(S): Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9b. OTHER REPORT NUMBER(S): If the report has been assigned any other report numbers *(either by the originator or by the sponsor)*, also enter this number(s).

10. AVAILABILITY/LIMITATION NOTICES: Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through _____."

(4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through _____."

(5) "All distribution of this report is controlled. Qualified DDC users shall request through _____."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. SUPPLEMENTARY NOTES: Use for additional explanatory notes.

12. SPONSORING MILITARY ACTIVITY: Enter the name of the departmental project office or laboratory sponsoring *(paying for)* the research and development. Include address.

13. ABSTRACT: Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as (TS), (S), (C), or (U).

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. KEY WORDS: Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, rules, and weights is optional.

UNCLASSIFIED
Security Classification

ANNEX V ITEM 2

DOCUMENT CONTROL DATA - R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)
Plasma Laboratory (R. W. Gould)
California Institute of Technology
Pasadena, California

2a. REPORT SECURITY CLASSIFICATION

2b GROUP

3. REPORT TITLE

Hydromagnetic Wave Boundary Condition and a Surface Wave in a Plasma Filled Waveguide

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)
Technical Report No. 2, July 1964

5. AUTHOR(S) (Last name, first name, initial)

SWANSON, D. G.

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
July 1964	26	13

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
Grant No. 412-63	TR No. 2
b. PROJECT NO.	
c.	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
d.	

10. AVAILABILITY/LIMITATION NOTICES

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
	Office of Aerospace Research A. F. Office of Scientific Research

13. ABSTRACT

The analysis of a magnetized plasma in a waveguide with a dielectric sheath between the plasma and waveguide is considered. Within the limitations of the cold plasma, effective dielectric tensor approach, the problem is solved exactly and a few illustrative computer solutions for the behavior of the transverse wave number are presented. Also, some approximate low frequency expressions are derived for the effect of the dielectric sheath. It is found that these solutions agree better with experiment than do those where no sheath at all is assumed, and appear adequate to account for all experimental observations. For the case of a finite or thick sheath, the solutions disagree with some other sheath theories, however, in an area where no experimental observations are yet reported.

The dielectric sheath also adds a surface wave to the group of hydromagnetic waves, and the coupling between the surface wave and the hydromagnetic waves is shown in certain frequency regions. Orthogonality relations are given which show that the surface wave and the hydromagnetic waves are all mutually orthogonal.

DD FORM 1 JAN 64 1473

UNCLASSIFIED
Security Classification

KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
surface wave sheath effect hydromagnetic waveguide plasma waves plasma-waveguide modes						

INSTRUCTIONS

1. ORIGINATING ACTIVITY: Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization *(corporate author)* issuing the report.

2a. REPORT SECURITY CLASSIFICATION: Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2b. GROUP: Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. REPORT TITLE: Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. DESCRIPTIVE NOTES: If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. AUTHOR(S): Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. REPORT DATE: Enter the date of the report as day, month, year, or month, year. If more than one date appears on the report, use date of publication.

7a. TOTAL NUMBER OF PAGES: The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7b. NUMBER OF REFERENCES: Enter the total number of references cited in the report.

8a. CONTRACT OR GRANT NUMBER: If appropriate, enter the applicable number of the contract or grant under which the report was written.

8b, 8c, & 8d. PROJECT NUMBER: Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9a. ORIGINATOR'S REPORT NUMBER(S): Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9b. OTHER REPORT NUMBER(S): If the report has been assigned any other report numbers *(either by the originator or by the sponsor)*, also enter this number(s).

10. AVAILABILITY/LIMITATION NOTICES: Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through _____."

(4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through _____."

(5) "All distribution of this report is controlled. Qualified DDC users shall request through _____."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. SUPPLEMENTARY NOTES: Use for additional explanatory notes.

12. SPONSORING MILITARY ACTIVITY: Enter the name of the departmental project office or laboratory sponsoring *(paying for)* the research and development. Include address.

13. ABSTRACT: Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as (TS), (S), (C), or (U).

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. KEY WORDS: Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, rules, and weights is optional.

UNCLASSIFIED

Security Classification

ANNEX V ITEM 3

DOCUMENT CONTROL DATA - R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)
Plasma Laboratory (R.W.Gould)
California Institute of Technology
Pasadena, California

2a. REPORT SECURITY CLASSIFICATION

2b. GROUP

3. REPORT TITLE

An Experimental Study of the Hydromagnetic Waveguide

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)
Technical Report No. 3

5. AUTHOR(S) (Last name, first name, initial)

HERTEL, Robert H.

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
January 1965	123	46

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
AF49(638)-1462	
b. PROJECT NO.	TR No. 3
c.	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
d.	

10. AVAILABILITY/LIMITATION NOTICES

Qualified requesters may obtain copies of this report from DDC.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
	Office of Aerospace Research A. F. Office of Scientific Research

13. ABSTRACT

The hydromagnetic waveguide consists of a cylindrical metal tube filled with a longitudinally magnetized plasma. Among the classes of waves which propagate in this system are the compressional hydromagnetic modes, characterized by a waveguide cutoff at low frequencies and by a resonance at the electron cyclotron frequency. This paper presents the results of observations of the propagation of such waves in a decaying hydrogen plasma at frequencies from 0.8 to 3.4 times the ion cyclotron frequency. The phase shift and attenuation of the waves are interpreted in terms of the ion density and the temperature by applying a theory based on a three-fluid description of the plasma. Spectroscopic measurements of the H_β line profile and absolute intensity are used to check the density and temperature inferred from the wave measurements.

The results of this study indicate that a simple approximate relationship between the phase factor and density obtained by neglecting dissipation gives densities which agree well with the spectroscopic measurements. As a diagnostic tool this method may yield densities to within ± 25% over a range of two decades. In the case of amplitude measurements only semiquantitative agreement between the wave and spectroscopic measurements is found, but the amplitude curves do show evidence of interferences between modes and a sharp cutoff at a critical density, both effects predicted by the theory.

DD FORM 1 JAN 64 1473

UNCLASSIFIED

Security Classification

UNCLASSIFIED

Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
hydromagnetic wave propagation(magnetohydrodynamic waves)						
plasma-filled waveguide						
plasma density measurement						
measurements in a decaying hydrogen plasma						
hydromagnetic wave interferometer						
plasma wave propagation						

INSTRUCTIONS

1. ORIGINATING ACTIVITY: Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization *(corporate author)* issuing the report.

2a. REPORT SECURITY CLASSIFICATION: Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2b. GROUP: Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. REPORT TITLE: Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. DESCRIPTIVE NOTES: If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. AUTHOR(S): Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. REPORT DATE: Enter the date of the report as day, month, year; or month, year. If more than one date appears on the report, use date of publication.

7a. TOTAL NUMBER OF PAGES: The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7b. NUMBER OF REFERENCES: Enter the total number of references cited in the report.

8a. CONTRACT OR GRANT NUMBER: If appropriate, enter the applicable number of the contract or grant under which the report was written.

8b, 8c, & 8d. PROJECT NUMBER: Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9a. ORIGINATOR'S REPORT NUMBER(S): Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9b. OTHER REPORT NUMBER(S): If the report has been assigned any other report numbers *(either by the originator or by the sponsor)*, also enter this number(s).

10. AVAILABILITY/LIMITATION NOTICES: Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through _____."

(4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through _____."

(5) "All distribution of this report is controlled. Qualified DDC users shall request through _____."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. SUPPLEMENTARY NOTES: Use for additional explanatory notes.

12. SPONSORING MILITARY ACTIVITY: Enter the name of the departmental project office or laboratory sponsoring *(paying for)* the research and development. Include address.

13. ABSTRACT: Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as (TS), (S), (C), or (U).

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. KEY WORDS: Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, rules, and weights is optional.

Security Classification

Unclassified
Security Classification

DOCUMENT CONTROL DATA - R&D
(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author) California Institute of Tech., Pasadena, Calif.	2a. REPORT SECURITY CLASSIFICATION Unclassified 2b. GROUP

3. REPORT TITLE

NONLINEAR EFFECTS IN TRAVELING WAVE LASER AMPLIFIERS

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)
Scientific Report

5. AUTHOR(S) (Last name, first name, initial)
Close, Donald H.

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
May, 1965	214	59

8a. CONTRACT OR GRANT NO. AF49(638)-1322 b. PROJECT NO. c. d.	9a. ORIGINATOR'S REPORT NUMBER(S) AFOSR 9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report) Scientific Report No. 5

10. AVAILABILITY/LIMITATION NOTICES
Agencies of the Department of Defense, their contractors and other Government agencies may obtain copies from DDC. All others apply to U. S. Dept. of Commerce, Office of Technical Services.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
~~Theoretical Study~~	Electronics Division Air Force Office of Scientific Research

13. ABSTRACT

Using semiclassical radiation theory, a formalism similar to that used by Lamb in his "Theory of an Optical Maser" is developed for studying the amplification of vector traveling waves in a laser-type medium. The effect of the medium on the waves is given in terms of space (or time) dependent field amplitudes and phases and a nonlinear index of refraction. With particular emphasis on typical gaseous media, the effects of Doppler broadening are treated in detail for arbitrary ratios of natural to Doppler line widths.

Lowest order nonlinear effects (due to a polarization cubic in the field amplitudes) are studied extensively, and the frequency dependence of several of these processes is presented in graphical form. The characteristics of these nonlinear processes peculiar to Doppler broadened lines are discussed, and the processes are interpreted in terms of saturation and coherent modulation of the population inversion density.

Strong nonlinear effects are considered in a more approximate way and are found to consist of saturation of the various linear and nonlinear processes previously considered. With the present formalism, the analytical results of Gordon, White and Rigden regarding gain saturation in laser amplifiers are obtained, and the extension is made to include frequencies away from line center and the effects of multiple spectral components. The introduction of fields at new frequencies is considered in detail. These results are also discussed in terms of saturation and coherent modulation of the populations and "hole burning".

DD FORM 1473

Unclassified

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Theory Nonlinear effects Index of Refraction Gain Saturation Frequency dependence Parametric effects Laser Doppler broadening						

INSTRUCTIONS

1. ORIGINATING ACTIVITY: Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization *(corporate author)* issuing the report.

2a. REPORT SECURITY CLASSIFICATION: Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2b. GROUP: Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. REPORT TITLE: Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. DESCRIPTIVE NOTES: If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. AUTHOR(S): Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. REPORT DATE: Enter the date of the report as day, month, year; or month, year. If more than one date appears on the report, use date of publication.

7a. TOTAL NUMBER OF PAGES: The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7b. NUMBER OF REFERENCES: Enter the total number of references cited in the report.

8a. CONTRACT OR GRANT NUMBER: If appropriate, enter the applicable number of the contract or grant under which the report was written.

8b, 8c, & 8d. PROJECT NUMBER: Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9a. ORIGINATOR'S REPORT NUMBER(S): Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9b. OTHER REPORT NUMBER(S): If the report has been assigned any other report numbers *(either by the originator or by the sponsor)*, also enter this number(s).

10. AVAILABILITY/LIMITATION NOTICES: Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through _____."

(4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through _____."

(5) "All distribution of this report is controlled. Qualified DDC users shall request through _____."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. SUPPLEMENTARY NOTES: Use for additional explanatory notes.

12. SPONSORING MILITARY ACTIVITY: Enter the name of the departmental project office or laboratory sponsoring *(paying for)* the research and development. Include address.

13. ABSTRACT: Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as *(TS)*, *(S)*, *(C)*, or *(U)*.

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. KEY WORDS: Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, rules, and weights is optional.

Indexing terms selected by indexers, author and by the FAST method (machine) for Sample Document TR No. 1

Term	Indexer #1	Indexer #2	Indexer #3	Indexer #4	Indexer #5	Indexer #6	Machine	Author
ATTENUATION	x	x	x		x		x	
BEAM			x					
CHARGE		x					x	
COLLISION	x	x	x		x		x	
COMPRESSION	x	x	x	x	x	x	x	
CUTOFF						x		
DAMPING								x
DECAY								x
DENSITY		x	x	x	x	x	x	x
DISPERSION		x	x				x	
ELECTROMAGNETICS		x						
EXCHANGE							x	
FIELD	x	x	x	x	x	x	x	x
FREQUENCY	x	x	x	x	x	x	x	x
FUNCTION	x	x					x	
GAS				x				
HYDROGEN	x	x	x	x	x	x	x	x

Term	Indexer #1	Indexer #2	Indexer #3	Indexer #4	Indexer #5	Indexer #6	Machine	Author
HYDROMAGNETICS	x	x	x	x		x	x	x
HYDROMAGNETISM					x			
IMPULSE		x	x	x	x		x	x
INTENSITY		x						
INTERATION		x						
ION		x	x	x			x	
IONIZATION	x	x	x	x	x	x	x	
LINEARITY							x	
MAGNETISM	x	x	x	x	x	x	x	
MAGNETOHYDRODYNAMIC								x
MASS				x			x	
MEASUREMENT					x		x	x
NEUTRAL			X					
PARTIAL		x						
PLASMA	x	x	x	x	x	x	x	x
PLASMA-FILLED					x			x
PROPAGATION	x	x	x		x		x	x
RANGE							x	
RESISTIVITY	x	x	x	x	x	x	x	x

Term	Indexer #1	Indexer #2	Indexer #3	Indexer #4	Indexer #5	Indexer #6	Machine	Author
RESPONSE				x	x		x	x
SPEED							x	
TEMPERATURE							x	
THEORY					x			
THERMAL							x	
TRANSFER		x	x				x	
TRANSFORMER	x						x	
VELOCITY		x						
WAVE	x	x	x	x	x	x	x	x
WAVEGUIDE	x	x	x	x	x	x	x	x

Indexing terms selected by indexers, author and by the FAST method (machine) for Sample Document TR No. 2.

Term	Indexer #1	Indexer #2	Indexer #3	Indexer #4	Indexer #5	Indexer #6	Machine	Author
ANALYSIS							x	
BEAM			x					
BEHAVIOR							x	
BOUNDARY	x	x		x	x	x	x	
COLD							x	
COMPUTER			x	x			x	
CONDITION	x			x				
COUPLING		x		x			x	
DIELECTRICS	x	x	x	x	x	x	x	
EFFECT								x
ELECTROMAGNETICS		x						
FREQUENCY		x			x		x	
GROUP							x	
HYDROMAGNETICS	x	x	x	x		x	x	x
HYDROMAGNETISM					x			
INDUCTION		x						
INTERACTION		x						
IONIZATION		x	x					

Term	Indexer #1	Indexer #2	Indexer #3	Indexer #4	Indexer #5	Indexer #6	Machine	Author
LOW					x			
MAGNETISM		x		x			x	
MAGNETOPLASMA			x					
MODE								x
OBSERVATION							x	
ORTHOGONALITY	x	x	x			x		
PARTIAL		x						
PLASMA	x	x	x	x	x	x	x	x
PLASMA-FILLED					x			
PLASMAGUIDE			x					
PLASMA-WAVEGUIDE								x
PROPAGATION	x							
RELATION			x					
SHEATH	x	x	x	x	x	x	x	x
SOLUTION			x	x				
SURFACE	x	x	x	x	x	x		x
TENSOR		x	x				x	
THICKNESS		x						
TRANSVERSE		x		x		x		
WAVE	x	x	x	x	x	x	x	x
WAVEGUIDE	x	x	x	x	x	x	x	x

Indexing terms selected by indexers, author and by the "AST method (machine) for Sample Document TR No. 3.

Term	Indexer #1	Indexer #2	Indexer #3	Indexer #4	Indexer #5	Indexer #6	Machine	Author
AMPLITUDE		x		x			x	
ATTENTUATION	x	x	x			x	x	
BEAM			x					
BETA				x				
CHARACTERISTIC							x	
COMPRESSION	x		x	x		x	x	
CRITICAL							x	
CURVE							x	
CYCLOTRON	x		x	x			x	
CYLINDER		x					x	
DECAYING					x		x	x
DENSITY	x	x	x	x	x	x	x	x
DIAGNOSIS							x	
DISSIPATION							x	
ELECTRON			x	x			x	
FLUID							x	
FREQUENCY	x	x	x	x			x	
HYDROGEN		x		x	x	x	x	x
HYDROMAGNETIC	x	x	x	x		x	x	x

Term	Indexer #1	Indexer #2	Indexer #3	Indexer #4	Indexer #5	Indexer #6	Machine	Author
HYDROMAGNETISM					x			
INDUCTION		x						
INTENSITY							x	
INTERFERENCE		x		x			x	
INTERFEROMETER								x
INTERPRETATION							x	
ION	x		x	x		x	x	
IONIZATION		x						
LINE				x				
LONGITUDINAL		x						
LOW			x					
MAGNETIC		x		x				
MAGNETISM					x		x	
MAGNETAHYDRODYNAMIC								x
MAGNETOPLASMA			x					
MEASUREMENT			x		x		x	x
METAL						x	x	
MODE		x						
OBSERVATION							x	

Term	Indexer #1	Indexer #2	Indexer #3	Indexer #4	Indexer #5	Indexer #6	Machine	Author
PAPER							x	
PHASE	x	x	x	x		x		
PLASMA	x	x		x	x	x	x	x
PLASMA-FILLED					x			x
PREDICTION							x	
PROFILE							x	
PROPAGATION	x	x	x	x	x		x	x
RANGE							x	
RESONANCE		x		x			x	
SHIFT	x	x	x			x		
SPECTROSCOPY	x	x	x	x	x		x	
TEMPERATURE	x	x	x		x	x	x	
THEORY					x			
TOOL							x	
TUBE						x	x	
WAVE	x	x	x	x	x	x	x	x
WAVEGUIDE	x	x	x	x	x	x	x	x

Indexing terms selected by indexers, author and by the FAST method (machine) for Sample Document TSR No. 5.

Term	Indexer #1	Indexer #2	Indexer #3	Indexer #4	Indexer #5	Indexer #6	Machine	Author
AMPLIFICATION	x	x		x	x		x	
AMPLIFIER			x	x	x	x	x	
AMPLITUDE	x			x			x	
APPROXIMATION	•						x	
ATOM		x						
BEAM			x					
BROADENING	x	x				x		x
CHARACTERISTIC							x	
COHERENCE	x	x	x		x	x	x	
DENSITY							x	
DEPENDENCY	x							x
DOPPLER	x	x		x		x		x
EFFECT	x	x	x	x		x		x
ELECTROMAGNETIC		x	x	x				
EXCITATION		x						
FIELD	x	x					x	
FORMALISM							x	

Term	Indexer #1	Indexer #2	Indexer #3	Indexer #4	Indexer #5	Indexer #6	Machine	Author
FREQUENCY	x	x	x	x			x	
GAIN		x			x			
GAS		x	x	x		x		
GRAPH							x	
HOLE BURNING					x			
INDEX	x			x	x		x	
INTERPRETATION							x	
INVERSION		x						
LASER	x	x	x	x	x	x	x	x
LIGHT		x		x				
LINE	x	x		x		x		
LINEAR			x				x	
MASER			x	x			x	
MEDIUM	x		x	x	x	x	x	
MODULATION	x	x	x	x	x	x	x	
NONLINEAR	x	x	x	x	x	x	x	x
OPTICAL				x				
OPTICS		x		x			x	
PHASE	x							
POLARIZATION							x	
POPULATION		x			x	x	x	
RADIATION	x	x	x	x			x	

Term	Indexer #1	Indexer #2	Indexer #3	Indexer #4	Indexer #5	Indexer #6	Machine	Author
REFRACTION	x			x	x		x	x
SATURATION	x	x	x		x	x	x	x
SPECTRUM	x	x					x	
THEORY	x				x			x
TIME	x						x	
TRANSITION		x						
TRAVEL	x			x				
TRAVELLING		x	x		x	x		
TUBE			x					
VECTOR					x		x	
WAVE	x	x	x	x	x	x	x	
PARAMETRIC EFFECTS								x
	30	24	26	18	23	17	13	15

ANNEX VII

Table of inter-indexer consistency coefficients for various combinations of indexers, authors, and machine for four sample documents. Numbers in the first column (indexer combination) designate an identifiable indexer, i.e. 1 stands for the indexer Joe Fix, 2 - for Mike Gibe, etc. M stands for machine (FAST Program) and A - for author. A set of numbers and/or letters indicate a particular indexer combination, for which the consistency coefficient was calculated. F. e., 1-3-6-M means indexes produced by indexers 1, 3, and 6, and by the FAST program (machine).

Indexer Combination	CONSISTENCY COEFFICIENT			
	Document TSR No. 1	Document TSR No. 2	Document TSR No. 3	Document TSR No. 5
1-2	0.555	0.409	0.440	0.428
1-3	0.608	0.421	0.700	0.312
1-4	0.500	0.562	0.521	0.468
1-5	0.565	0.466	0.318	0.366
1-6	0.666	0.750	0.611	0.392
2-3	0.678	0.384	0.366	0.375
2-4	0.482	0.478	0.500	0.400
2-5	0.483	0.347	0.285	0.303
2-6	0.444	0.500	0.384	0.464
3-4	0.583	0.428	0.444	0.413
3-5	0.576	0.285	0.269	0.346
3-6	0.545	0.444	0.269	0.500
4-5	0.541	0.388	0.259	0.290
4-6	0.666	0.600	0.360	0.357
5-6	0.500	0.500	0.272	0.454
1-M	0.516	0.333	0.325	0.421
2-M	0.583	0.423	0.333	0.317
3-M	0.575	0.320	0.325	0.371
4-M	0.500	0.476	0.380	0.394
5-M	0.500	0.333	0.268	0.424
6-M	0.375	0.350	0.300	0.285

1-A	0.291	0.428	0.285	0.423
2-A	0.281	0.260	0.259	0.258
3-A	0.333	0.315	0.240	0.192
4-A	0.958	0.352	0.291	0.250
5-A	0.458	0.333	0.529	0.304
6-A	0.333	0.461	0.300	0.272
M-A	0.314	0.238	0.219	0.162
1-2-3	0.583	0.285	0.366	0.225
1-2-4	0.366	0.333	0.321	0.256
1-2-5	0.406	0.280	0.225	0.184
1-2-6	0.392	0.409	0.333	0.278
1-3-4	0.423	0.318	0.407	0.216
1-3-5	0.464	0.260	0.230	0.194
1-3-6	0.458	0.400	0.434	0.242
1-4-5	0.384	0.350	0.206	0.216
1-4-6	0.478	0.500	0.320	0.235
1-5-6	0.416	0.437	0.208	0.226
2-3-4	0.451	0.259	0.250	0.237
2-3-5	0.454	0.206	0.171	0.189
2-3-6	0.413	0.307	0.187	0.281
2-4-5	0.363	0.269	0.205	0.132
2-4-6	0.400	0.391	0.233	0.228
2-5-6	0.343	0.304	0.187	0.242
3-4-5	0.428	0.240	0.156	0.176
3-4-6	0.480	0.333	0.241	0.258
3-5-6	0.407	0.272	0.142	0.310
4-5-6	0.440	0.368	0.161	0.187
1-2-M	0.416	0.250	0.222	0.239
1-3-M	0.424	0.222	0.279	0.214
1-4-M	0.366	0.304	0.255	0.262
1-5-M	0.382	0.250	0.162	0.250
1-6-M	0.343	0.318	0.225	0.179
2-3-M	0.473	0.225	0.187	0.204
2-4-M	0.378	0.321	0.255	0.182
2-5-M	0.410	0.241	0.166	0.204
2-6-M	0.324	0.269	0.177	0.195
3-4-M	0.411	0.259	0.239	0.225
3-5-M	0.416	0.178	0.152	0.237
3-6-M	0.352	0.230	0.186	0.237
4-5-M	0.371	0.250	0.155	0.220
4-6-M	0.363	0.318	0.186	0.154
5-6-M	0.314	0.260	0.139	0.263
1-2-A	0.212	0.240	0.200	0.194
1-3-A	0.241	0.272	0.200	0.147
1-4-A	0.222	0.315	0.214	0.206
1-5-A	0.230	0.277	0.208	0.194
1-6-A	0.240	0.400	0.217	0.200
2-3-A	0.264	0.206	0.147	0.135

2-4-A	0.235	0.230	0.212	0.132
2-5-A	0.235	0.192	0.200	0.114
2-6-A	0.181	0.260	0.193	0.182
3-4-A	0.275	0.250	0.161	0.118
3-5-A	0.275	0.208	0.178	0.100
3-6-A	0.250	0.285	0.148	0.143
4-5-A	0.296	0.238	0.206	0.118
4-6-A	0.291	0.333	0.200	0.125
5-6-A	0.240	0.294	0.208	0.120
1-2-3-4	0.343	0.250	0.250	0.167
1-2-3-5	0.382	0.193	0.171	0.143
1-2-3-6	0.366	0.285	0.250	0.175
1-2-4-5	0.294	0.259	0.176	0.122
1-2-4-6	0.354	0.333	0.200	0.179
1-2-5-6	0.303	0.280	0.151	0.158
1-3-4-5	0.333	0.230	0.156	0.125
1-3-4-6	0.407	0.318	0.241	0.162
1-3-5-6	0.344	0.250	0.142	0.194
1-4-5-6	0.370	0.350	0.129	0.135
2-3-4-5	0.342	0.200	0.135	0.095
2-3-4-6	0.375	0.259	0.147	0.158
2-3-5-6	0.323	0.206	0.108	0.189
2-4-5-6	0.323	0.269	0.138	0.105
3-4-5-6	0.379	0.240	0.088	0.171
1-2-3-M	0.378	0.181	0.187	0.167
1-2-4-M	0.297	0.241	0.170	0.149
1-2-5-M	0.333	0.193	0.145	0.149
1-2-6-M	0.306	0.250	0.155	0.130
1-3-4-M	0.323	0.214	0.217	0.163
1-3-5-M	0.361	0.200	0.130	0.159
1-3-6-M	0.323	0.214	0.186	0.163
1-4-5-M	0.285	0.230	0.130	0.178
1-4-6-M	0.333	0.304	0.162	0.119
1-5-6-M	0.285	0.240	0.116	0.171
2-3-4-M	0.368	0.187	0.140	0.130
2-3-5-M	0.365	0.147	0.117	0.152
2-3-6-M	0.307	0.193	0.125	0.159
2-4-5-M	0.300	0.193	0.140	0.109
2-4-6-M	0.315	0.250	0.135	0.091
2-5-6-M	0.275	0.206	0.147	0.182
3-4-5-M	0.324	0.166	0.108	0.140
3-4-6-M	0.342	0.222	0.111	0.146
3-5-6-M	0.297	0.172	0.088	0.220
4-5-6-M	0.305	0.240	0.077	0.143
1-2-3-A	0.200	0.193	0.147	0.122
1-2-4-A	0.171	0.222	0.181	0.125
1-2-5-A	0.171	0.178	0.151	0.077
1-2-6-A	0.176	0.240	0.156	0.162

1-3-4-A	0.193	0.240	0.161	0.105
1-3-5-A	0.193	0.192	0.142	0.079
1-3-6-A	0.200	0.260	0.148	0.114
1-4-5-A	0.172	0.217	0.161	0.102
1-4-6-A	0.214	0.315	0.166	0.111
1-5-6-A	0.185	0.263	0.153	0.091
2-3-4-A	0.222	0.200	0.138	0.098
2-3-5-A	0.222	0.156	0.108	0.077
2-3-6-A	0.200	0.206	0.111	0.108
2-4-5-A	0.194	0.172	0.166	0.050
2-4-6-A	0.200	0.230	0.171	0.105
2-5-6-A	0.171	0.192	0.147	0.086
3-4-5-A	0.225	0.178	0.117	0.054
3-4-6-A	0.233	0.250	0.121	0.086
3-5-6-A	0.200	0.200	0.100	0.094
4-5-6-A	0.214	0.227	0.151	0.059
1-2-3-4-5	0.277	0.193	0.135	0.091
1-2-3-4-6	0.333	0.250	0.147	0.119
1-2-3-5-6	0.285	0.193	0.108	0.143
1-2-4-5-6	0.285	0.259	0.111	0.098
1-3-4-5-6	0.333	0.230	0.088	0.125
2-3-4-5-6	0.305	0.200	0.078	0.098
1-2-3-4-M	0.307	0.181	0.140	0.122
1-2-3-5-M	0.317	0.138	0.120	0.122
1-2-3-6-M	0.282	0.181	0.125	0.125
1-2-4-5-M	0.250	0.187	0.120	0.104
1-2-4-6-M	0.289	0.241	0.106	0.085
1-2-5-6-M	0.250	0.193	0.104	0.125
1-3-4-5-M	0.270	0.161	0.102	0.109
1-3-4-6-M	0.314	0.214	0.130	0.114
1-3-5-6-M	0.270	0.161	0.087	0.159
1-4-5-6-M	0.277	0.230	0.087	0.111
2-3-4-5-M	0.285	0.142	0.094	0.083
2-3-4-6-M	0.300	0.187	0.080	0.087
2-3-5-6-M	0.261	0.147	0.078	0.152
2-4-5-6-M	0.268	0.193	0.100	0.087
3-4-5-6-M	0.289	0.166	0.061	0.136
1-2-3-4-A	0.162	0.193	0.138	0.093
1-2-3-5-A	0.162	0.147	0.108	0.068
1-2-3-6-A	0.166	0.193	0.111	0.098
1-2-4-5-A	0.135	0.166	0.138	0.048
1-2-4-6-A	0.166	0.222	0.142	0.100
1-2-5-6-A	0.138	0.178	0.114	0.077
1-3-4-5-A	0.151	0.172	0.117	0.049
1-3-4-6-A	0.187	0.240	0.121	0.077
1-3-5-6-A	0.156	0.185	0.100	0.081
1-4-5-6-A	0.166	0.217	0.121	0.053
2-3-4-5-A	0.184	0.151	0.102	0.046

2-3-4-6-A	0.189	0.200	0.105	0.073
2-3-5-6-A	0.162	0.156	0.077	0.077
2-4-5-6-A	0.162	0.172	0.131	0.050
3-4-5-6-A	0.187	0.178	0.083	0.054
1-2-3-4-5-6	0.270	0.193	0.077	0.091
1-2-3-4-5-M	0.238	0.138	0.094	0.080
1-2-3-4-6-M	0.275	0.181	0.080	0.082
1-2-3-5-6-M	0.238	0.138	0.078	0.122
1-2-4-5-6-M	0.243	0.187	0.080	0.083
1-3-4-5-6-M	0.263	0.161	0.061	0.109
2-3-4-5-6-M	0.255	0.142	0.057	0.083
1-2-3-4-5-A	0.128	0.147	0.102	0.044
1-2-3-4-6-A	0.157	0.193	0.105	0.070
1-2-3-5-6-A	0.131	0.147	0.077	0.070
1-2-4-5-6-A	0.166	0.166	0.105	0.048
1-3-4-5-6-A	0.147	0.172	0.083	0.049
2-3-4-5-6-A	0.153	0.151	0.073	0.046

Indexing terms assigned by human indexers and by the FAST program to sample document TSR No. 1 in intra-indexing consistency test.

Term	Indexer #1	Indexer #4	Indexer #5	Indexer #6	Machine (FAST)
ATTENUATION	⊗	○	x		⊗
CHARGE					⊗
COLLISION	⊗	○	x	○	⊗
COMPRESSION	⊗	⊗	⊗	⊗	⊗
CUTOFF			○	x	
DENSITY	○	⊗	x	x	⊗
DISPERSION					⊗
EXCHANGE					⊗
FIELD	⊗	⊗	⊗	⊗	⊗
FREQUENCY	⊗	⊗	⊗	⊗	⊗
FUNCTION	⊗	○	○		⊗
HYDROGEN	⊗	⊗	⊗	⊗	⊗
HYDROMAGNETICS	⊗	⊗		⊗	⊗
HYDROMAGNETISM			⊗		
IMPULSE		x	x		⊗
ION	○	⊗		○	⊗
IONIZATION	⊗	⊗	⊗	⊗	⊗
LINEARITY					⊗

Term	Indexer #1	Indexer #4	Indexer #5	Indexer #6	Machine (FAST)
MAGNETISM	⊗	⊗	⊗	⊗	⊗
MASS		⊗			⊗
MEASUREMENT			x		⊗
NEUTRAL	○			○	
PLASMA	⊗	⊗	⊗	⊗	⊗
PLASMA-FILLED			x		
PROPAGATION	⊗	○	x		⊗
RANGE					⊗
RESISTIVITY	⊗	x	⊗	⊗	⊗
RESPONSE		x	x		⊗
SPEED					⊗
TEMPERATURE					⊗
THEORY			⊗		
THERMAL					⊗
TRANSFER	○	○	○		⊗
TRANSFORMER	x				⊗
WAVE	⊗	⊗	⊗	⊗	⊗
WAVEGUIDE	⊗	⊗	⊗	⊗	⊗
COLD			○		
CONSISTENCY	75%	59.1%	50%	68.7%	100%

Indexing terms assigned by human indexers and by the FAST program to sample document TSR No. 2 in intra-indexing consistency test.

Term	Indexer #1	Indexer #4	Indexer #5	Indexer #6	Machine (FAST)
ANALYSIS		○	○		⊗
BEHAVIOR					⊗
BOUNDARY	⊗	⊗	⊗	⊗	⊗
COLD			○		⊗
COMPUTER		⊗	○		⊗
CONDITION	x	x		○	
COUPLING		⊗			⊗
DIELECTRICS	⊗	⊗	⊗	⊗	⊗
FREQUENCY	○		⊗		⊗
GROUP					⊗
HYDROMAGNETICS	⊗	⊗		⊗	⊗
HYDROMAGNETISM			⊗		
LOW			⊗		
MAGNETISM		⊗	○		⊗
OBSERVATION					⊗
ORTHOGONALITY	⊗			⊗	

Term	Indexer #1	Indexer #4	Indexer #5	Indexer #6	Machine (FAST)
PLASMA	⊗	⊗	⊗	⊗	⊗
PLASMA-FILLED			x		
PROPAGATION	x				
RELATION				○	
SHEATH	⊗	⊗	⊗	⊗	⊗
SOLUTION		x			
SURFACE	⊗	⊗	⊗	⊗	
TENSOR	○	○			⊗
TRANSVERSE	○	⊗		x	
WAVE	⊗	⊗	⊗	⊗	⊗
WAVEGUIDE	⊗	⊗	⊗	⊗	⊗
NUMBER		○			
MAGNETIZED				○	
CONSISTENCY	64.3%	70.6%	66.6%	75.0%	100.0%

Indexing terms assigned by human indexers and by the FAST program to sample document TSR No. 3 in intra-indexing consistency test.

Term	Indexer #1	Indexer #4	Indexer #5	Indexer #6	Machine
AMPLITUDE		⊗	○		⊗
ATTENTUATION	⊗	○	○	x	⊗
BETA		x			
CHARACTERISTIC					⊗
COMPRESSION	x	⊗	○	⊗	⊗
CRITICAL					⊗
CURVE					⊗
CYCLOTRON	⊗	⊗			⊗
CYLINDER					⊗
DECAYING			⊗		⊗
DENSITY	⊗	⊗	⊗	⊗	⊗
DIAGNOSIS					⊗
DISSIPATION					⊗
ELECTRON		x			
FLUID					⊗
FREQUENCY	⊗	⊗			⊗
HYDROGEN	○	⊗	⊗	⊗	⊗
HYDROMAGNETIC	⊗	⊗	⊗	⊗	⊗

Term	Indexer #1	Indexer #4	Indexer #5	Indexer #6	Machine
INTENSITY					⊗
INTERFERENCE		×			⊗
INTERPRETATION					⊗
ION	⊗	⊗	○	⊗	⊗
LINE		×			
MAGNETISM	○	⊗	×		⊗
MEASUREMENT			⊗	○	⊗
METAL				×	⊗
OBSERVATION					⊗
PAPER					⊗
PHASE	⊗	⊗	○	×	
PLASMA	⊗	⊗	⊗	⊗	⊗
PLASMA-FILLED			×		
PREDICTION					⊗
PROFILE					⊗
PROPAGATION	⊗	⊗	⊗	○	⊗
RANGE					⊗
RESONANCE		×			⊗
SHIFT	⊗	○	○	×	
SPECTROSCOPY	⊗	⊗	⊗	○	⊗
TEMPERATURE	×		⊗	⊗	⊗
THEORY			⊗		

Term	Indexer #1	Indexer #4	Indexer #5	Indexer #6	Machine
TOOL					⊗
TUBE				x	⊗
WAVE	⊗	⊗	⊗	⊗	⊗
WAVEGUIDE	⊗	⊗	⊗	⊗	⊗
CORRELATION		○			
MAGNETIZED				○	
CONSISTENCY	76.5%	65.2%	60.0%	52.9%	100.0%

Indexing terms assigned by human indexers and by the FAST program to sample document TSR No. 5 in intra-indexing consistency test.

Term	Indexer #1	Indexer #4	Indexer #5	Indexer #6	Machine
AMPLIFICATION	x	x	⊗		⊗
AMPLIFIER	○	⊗	⊗	⊗	⊗
AMPLITUDE	⊗	⊗			⊗
APPROXIMATION					⊗
BROADENING	x		○	⊗	
CHARACTERISTIC					⊗
COHERENCE	⊗	○	⊗	x	⊗
DENSITY			○		⊗
DEPENDENCE	⊗				
DOPPLER	⊗	⊗	○	⊗	
EFFECT	⊗	⊗		⊗	
ELECTROMAGNETIC		x			
FIELD	⊗	○			⊗
FORMALISM					⊗
FREQUENCY	⊗	⊗			⊗
GAIN		○	⊗		
GAS		⊗		⊗	
GRAPH					⊗

Term	Indexer #1	Indexer #4	Indexer #5	Indexer #6	Machine
HOLE-BURNING			⊗		
INDEX	⊗	×	×		⊗
INTERPRETATION					⊗
INVERSION		○	○		
LASER	⊗	⊗	⊗	⊗	⊗
LIGHT		×			
LINE	×	×	○	⊗	
LINEAR					⊗
MASER		⊗			⊗
MEDIUM	×	⊗	×	⊗	⊗
MODULATION	⊗	⊗	⊗	⊗	⊗
NONLINEAR	⊗	⊗	⊗	⊗	⊗
OPTICAL		×			
OPTICS		⊗			⊗
PHASE	⊗				
POLARIZATION	○	⊗			⊗
POPULATION			⊗	⊗	⊗
RADIATION	×	⊗			⊗
REFRACTION	⊗	×	×		⊗
SATURATION	⊗	○	⊗	⊗	⊗

Term	Indexer #1	Indexer #4	Indexer #5	Indexer #6	Machine
SPECTRUM	⊗				⊗
THEORY	x		⊗		
TIME	⊗				⊗
TRAVEL	⊗	⊗			
TRAVELLING			⊗	⊗	⊗
VECTOR	○		⊗		⊗
WAVE	⊗	⊗	⊗	⊗	⊗
SPACE	○				
CONSISTENCY	64.2%	57.1%	59.0%	93.3%	100.0%

ANNEX IX

Frequency distribution of terms by number of postings for ILSE 1963 index.

No. of Postings (u_i)	No. of Terms $F(u_i)$	Relative Frequency of Terms with u_i Postings $f(u_i)$
1	1342	0.4266
2	462	0.1468
3	268	0.0852
4	156	0.0496
5	111	0.0353
6	75	0.0238
7	72	0.0229
8	61	0.0194
9	37	0.0118
10	37	0.0118
11	29	0.0092
12	35	0.0111
13	24	0.0076
14	23	0.0073
15	20	0.0063
16	20	0.0063
17	17	0.0054
18	11	0.0034
19	2	0.0067
20	22	0.0070
21	13	0.0041
22	8	0.0025
23	14	0.0044
24	8	0.0025
25	7	0.0022
26	11	0.0034
27	9	0.0028
28	9	0.0028
29	6	0.0019
30	8	0.0025
31	8	0.0025
32	4	0.0012
33	4	0.0012
34	9	0.0028
35	4	0.0012
36	3	0.0009
37	2	0.0006

No. of Postings (u_i)	No. of Terms $F(u_i)$	Relative Frequency of Terms with u_i Postings $f(u_i)$
38	1	0.0003
39	4	0.0012
40	4	0.0012
41	5	0.0015
42	5	0.0015
43	2	0.0006
44	4	0.0012
45	3	0.0009
46	4	0.0012
47	3	0.0009
48	3	0.0009
49	2	0.0006
50	4	0.0012
51	2	0.0006
52	2	0.0006
54	1	0.0003
55	2	0.0006
56	2	0.0006
57	2	0.0006
58	1	0.0003
59	2	0.0006
60	2	0.0006
61	2	0.0006
62	2	0.0006
64	2	0.0006
65	1	0.0003
66	2	0.0006
67	4	0.0012
68	1	0.0003
69	2	0.0006
70	2	0.0006
71	5	0.0015
72	2	0.0006
73	5	0.0015
74	2	0.0006
76	1	0.0003
78	1	0.0003
79	1	0.0003
80	1	0.0003
81	1	0.0003
82	1	0.0003
83	2	0.0006

No. of Postings (u_i)	No. of Terms $F(u_i)$	Relative Frequency of Terms with u_i Postings $f(u_i)$
84	2	0.0006
85	2	0.0006
86	2	0.0006
88	1	0.0003
89	2	0.0006
91	1	0.0003
94	2	0.0006
96	1	0.0003
97	1	0.0003
98	1	0.0003
99	1	0.0003
100	1	0.0003
102	1	0.0003
103	2	0.0006
105	1	0.0003
107	1	0.0003
109	1	0.0003
114	2	0.0006
116	1	0.0003
118	1	0.0003
121	1	0.0003
122	1	0.0003
124	1	0.0003
125	1	0.0003
127	1	0.0003
129	1	0.0003
130	1	0.0003
132	2	0.0006
134	1	0.0003
135	1	0.0003
137	2	0.0006
138	1	0.0003
144	1	0.0003
146	1	0.0003
158	1	0.0003
162	1	0.0003
163	1	0.0003
165	1	0.0003
166	1	0.0003
170	1	0.0003
171	1	0.0003
172	1	0.0003

No. of Postings (u_i)	No. of Terms $F(u_i)$	Relative Frequency of Terms with u_i Postings $f(u_i)$
191	1	0.0003
194	1	0.0003
197	1	0.0003
218	1	0.0003
221	1	0.0003
248	1	0.0003
250	1	0.0003
267	1	0.0003
277	1	0.0003
288	1	0.0003
295	1	0.0003
316	1	0.0003
317	1	0.0003
322	1	0.0003
380	1	0.0003
388	1	0.0003
396	1	0.0003
426	1	0.0003
569	1	0.0003
710	1	0.0003
1,065	1	0.0003
1,072	1	0.0003
1,251	1	0.0003
1,310	1	0.0003
1,381	1	0.0003
Σ = 37,471	Σ = 3,146	

ANNEX XI

REPORT TITLE

HYDROMAGNETIC WAVE BOUNDARY CONDITION AND A SURFACE WAVE IN A PLASMA FILLED WAVEGUIDE

AUTHOR(S)

SWANSON, D. G.

ABSTRACT

THE ANALYSIS OF A MAGNETIZED PLASMA IN A WAVEGUIDE WITH A DIELECTRIC SHEATH BETWEEN THE PLASMA AND WAVEGUIDE IS CONSIDERED. WITHIN THE LIMITATIONS OF THE COLD PLASMA, EFFECTIVE DIELECTRIC TENSOR APPROACH, THE PROBLEM IS SOLVED EXACTLY AND A FEW ILLUSTRATIVE COMPUTER SOLUTIONS FOR THE BEHAVIOR OF THE TRANSVERSE WAVE NUMBER ARE PRESENTED. ALSO, SOME APPROXIMATE LOW FREQUENCY EXPRESSIONS ARE DERIVED FOR THE EFFECT OF THE DIELECTRIC SHEATH. IT IS FOUND THAT THESE SOLUTIONS AGREE BETTER WITH EXPERIMENT THAN DO THOSE WHERE NO SHEATH AT ALL IS ASSUMED, AND APPEAR ADEQUATE TO ACCOUNT FOR ALL EXPERIMENTAL OBSERVATIONS. FOR THE CASE OF A FINITE OR THICK SHEATH, THE SOLUTIONS DISAGREE WITH SOME OTHER SHEATH THEORIES, HOWEVER, IN AN AREA WHERE NO EXPERIMENTAL OBSERVATIONS ARE YET REPORTED.

THE DIELECTRIC SHEATH ALSO ADDS A SURFACE WAVE TO THE GROUP OF HYDROMAGNETIC WAVES, AND THE COUPLING BETWEEN THE SURFACE WAVE AND THE HYDROMAGNETIC WAVES IS SHOWN IN CERTAIN FREQUENCY REGIONS. ORTHOGONALITY RELATIONS ARE GIVEN WHICH SHOW THAT THE SURFACE WAVE AND THE HYDROMAGNETIC WAVES ARE ALL MUTUALLY ORTHOGONAL.

KEY WORDS: ANALYSIS, BEHAVIOR, BOUNDARY, COLD, COMPUTER, COUPLING, DIELECTRICS, FREQUENCY, GROUP, HYDROMAGNETICS, MAGNETISM, OBSERVATIONS, PLASMA, SHEATH, TENSOR, WAVE, WAVEGUIDE.

BIBLIOGRAPHY

A. AUTOMATIC INDEXING

1. Arnovick, G. N., Liles, J. A., and Wood, J. S. Information Storage and Retrieval: Analysis of the State-of-the-Art. AFIPS Conference Proceedings: Spring Joint Computer Conference, Washington, D. C., 1964, Vol. 25. Spartan Books, Baltimore, Md., 1964, pp. 537-561.

2. Artandi, S. A Selective Bibliographic Survey of Automatic Indexing Methods. Special Libraries, Vol. 54, December 1963, pp. 630-634.

3. Artandi, S. Indexing by Computer. (Unpublished PhD. Thesis) Rutgers University, Graduate School of Library Service, New Brunswick, N. J., 1963, p. 200.

4. Artandi, S. Mechanical Indexing of Proper Nouns. Journal of Documentation, Vol. 19, No. 4, December 1963, pp. 187-196.

5. Artandi, S. Thesaurus Controls Automatic Book Indexing by Computers. Automatic and Scientific Communications. ADI 26th Annual Meeting, Chicago, Illinois, October 7-11, 1963. American Documentation Institute, Washington, D. C., 1963, pp. 1-2.

6. Artandi, S. Automatic Book Indexing by Computer. American Documentation, Vol. 15, No. 4, October 1964, pp. 250-257.

7. Baxendale, P. B. Machine-Made Index for Technical Literature: An Experiment. IBM Journal of Research and Development, Vol. 2, No. 4, October 1958, pp. 354-361.

8. Baxendale, P. B. An Empirical Model for Machine Indexing. Institute on Information Storage and Retrieval. Machine Indexing: Progress and Problems. Papers Presented at the 3rd Institute, February 13-17, 1961. The American University, Washington, D. C., 1962, pp. 207-218.

9. Baxendale, P. B. Automatic Processing for a Limited Type of Document Retrieval System. Automation and Scientific Communications: ADI 26th Annual Meeting, Chicago, Illinois, October 7-11, 1963. American Documentation Institute, Washington, D. C., 1963, pp. 67-68.

10. Borko, H. Automatic Document Classification Using a Mathematically Derived Classification System. System Development Corporation, Santa Monica, Calif., December 1961, p. 5. AD 282 539.

11. Borko, H. The Construction of an Empirically-Based Mathematically-Derived Classification System. AFIPS Conference Proceedings: Spring Joint Computer Conference, San Francisco, Calif., May 1-3, 1962, Vol. 21. National Press, Palo Alto, Calif., 1962, pp. 279-287.

12. Borko, H. Measuring the Reliability of Subject Classification by Men and Machines. American Documentation, Vol. 15, No. 4, 1964.

13. Borko, H. Research in Automatic Generation of Classification Systems. AFIPS Conference Proceedings: Spring Joint Computer Conference, Washington, D. C., 1964, Vol. 25. Spartan Books, Baltimore, Md., 1964, pp. 529-535.

14. Borko, H. and Bernick, M. Automatic Document Classification. Journal of the Association for Computing Machinery, Vol. 10, No. 2, 1963, pp. 151-162.

15. Borko, H. and Bernick, M. Automatic Document Classification. Part II. Additional Experiments. Journal of the Association for Computing Machinery, Vol. 11, No. 2, April 1964, pp. 138-151.

16. Brandenburg, W., Fallon, H. C., Hensley, C. B., Savage, T. R., and Sowarby, A. J. Selective Dissemination of Information. SDI 2 System Report 17-031. International Business Machines Corporation, Yorktown Heights, N. Y., April 1961.

17. Climenson, W. D., et al. Automatic Syntax Analysis in Machine Indexing and Abstracting. American Documentation, Vol. 12, July 1961, pp. 178-183.

18. Climenson, W. D., Hardwick, N. H., and Jacobson, S. N. Automatic Syntax Analysis in Machine Indexing and Abstracting. Institute on Information Storage and Retrieval. Machine Indexing: Progress and Problems. Papers Presented at the 3rd Institute, February 13-17, 1961. The American University, Washington, D. C., 1962, pp.305-325.

19. Cornelius, M. E. Machine Input Problems for Machine Indexing, Alternatives and Practicalities. Paper Presented for the 3rd Institute on Information Storage and Retrieval, American University, Washington, D. C., February 13, 1961, p. 8.

20. Cornelius, M. E. Machine Input for Machine Indexing: Alternatives and Practicalities. Institute on Information Storage and Retrieval. Machine Indexing: Progress and Problems. Papers Presented at the 3rd Institute, February 13-17, 1961. The American University, Washington, D. C., 1962, pp. 41-49.

21. Claridge, P. R. P. Mechanized Indexing of Information on Chemical Compounds in Plants. The Indexer, Vol. 2, No. 1, Spring 1960, pp. 4-19.

22. Edmundson, H. P., Oswald, V. A., Jr., and Wyllys, R. E. Automatic Indexing and Abstracting of the Contents of Documents. Planning Research Corporation, Los Angeles, Calif., 1959, p. 133. AD 231 606.

23. Edmundson, H. P. and Wyllys, R. E. Automatic Abstracting and Indexing. Survey and Recommendations. Journal of the Association for Computing Machinery, Vol. 4, No. 5, May 1961, pp. 226-234.

24. Feldman, A., Holland, D. B., and Jacobus, D. P. The Automatic Encoding of Chemical Structures. Journal of Chemical Documentation, Vol. 3, No. 4, October 1963, pp. 187-188.

25. Heald, J. H. Transition from a Manual to a Machine Indexing System. Institute on Information Storage and Retrieval. Machine Indexing: Progress and Problems. Papers Presented at the 3rd Institute, February 13-17, 1961. The American University, Washington, D. C., 1962, pp. 170-190.

26. Hillman, D. J. Mathematical Theories of Revelance with Respect to Systems of Automatic and Manual Indexing. Automation and Scientific Communication: ADI 26th Annual Meeting, Chicago, Illinois, October 7-11, 1963. American Documentation Institute, Washington, D. C., 1963, pp. 323-324.

27. Howerton, P. W. The Application of Modern Leciographic Techniques to Machine Indexing. Institute on Information Storage and Retrieval. Machine Indexing: Progress and Problems. Papers Presented at the 3rd Institute, February 13-17, 1961. The American University, Washington, D. C., 1962, pp. 326-330.

28. Kraft, D. H. An Operational Selective Dissemination of Information (SDI) System for Technical and Non-Technical Personnel Using Automatic Indexing Techniques. Automatic and Scientific Communication: ADI 26th Annual Meeting, Chicago, Illinois, October 7-11, 1963. American Documentation Institute, Washington, D. C., 1963, pp. 69-70.

29. Kurmey, W. J. An Evaluation of Automatically Prepared Abstracts and Indexes, a Dissertation. The University of Chicago, Chicago, Illinois, 1964.

30. Levery F. An Experiment in Automatic Indexing of French Language Documents. Automation and Scientific Communication: American Documentation Institute, 26th Annual Meeting, Chicago, Illinois, October 7-11, 1963, pp. 235-236.

31. Luhn, H. P. A Statistical Approach to Mechanized Encoding and Searching of Literary Information. IBM Journal of Research and Development, Vol. 1, No. 4, October 1957, pp. 309-317.

32. Luhn, H. P. The Automatic Creation of Literature Abstracts. IBM Journal of Research and Development, Vol. 2, No. 2, April 1958, pp. 159-165.

33. Luhn, H. P. Autoencoding of Documents for Information Retrieval Systems. International Business Machines Corporation, Yorktown Heights, N. Y., 1958.

34. Luhn, H. P. Potentialities of Autoencoding of Scientific Literature. IBM Research Report RC-101. International Business Machines Corporation, Yorktown Heights, N. Y., May 15, 1959, p. 22.

35. Luhn, H. P. The Automatic Derivation of Information Retrieval Encodements from Machine-Readable Texts. International Business Machines Corporation, Yorktown Heights, N. Y., 1959.

36. Maron, M. E. Automatic Indexing: an Experimental Inquiry. The Rand Corporation, Santa Monica, Calif., 1960, p. 31. AD 245 175

37. Maron, M. E. Automatic Indexing: an Experimental Inquiry. Journal of the Association for Computing Machinery, Vol. 8, No. 3, July 1961, pp. 404-417.

38. Maron, M. E. Automatic Indexing: an Experimental Inquiry. Institute on Information Storage and Retrieval. Machine Processing: Progress and Problems. Papers Presented at the 3rd Institute, February 13-17, 1961. The American University, Washington, D. C., 1962, pp. 236-264.

39. Melton, J. S., Reeves, P. W., and Hespen, C. F., Automatic Processing of Metallurgical Abstracts for the Purpose of Information Retrieval, Center for Documentation and Communication Research, Western Reserve University, Cleveland, Ohio, February 1964, p. 102.

40. O'Connor, J. Mechanized Indexing: Some General Remarks and Some Small-Scale Empirical Results. Institute for Cooperative Research, University of Pennsylvania, Philadelphia, Pa., 1960, p. 57. AD 250 209, PB 159 398.

41. O'Connor, J. Some Suggested Mechanized Indexing Investigations Which Require No Machines. Institute for Cooperative Research, University of Pennsylvania, Philadelphia, Pa., 1960, p. 19. AD 240 040.

42. O'Connor, J. Some Remarks on Mechanized Indexing and Small-Scale Empirical Results. Institute on Information Storage and Retrieval. Machine Indexing: Progress and Problems. Papers Presented at the 3rd Institute, February 13-17, 1961. The American University, Washington, D. C., 1962, pp. 266-279.

43. O'Connor, J. Methods of Mechanized Indexing: Comprehensive Document Preparation and Testing Mechanized Indexing Quality. Institute for Cooperation Research, University of Pennsylvania, Philadelphia, Pa., 1963, p. 29. AD 412 723.

44. O'Connor, J. Mechanized Indexing Methods and Their Testing. Institute for Scientific Information, Philadelphia, Pa., 1963, p. 29. AD 409 276.

45. O'Connor, J. Mechanized Indexing Studies of MSD Toxicity. Institute for Scientific Information, Philadelphia, Pa., January 1964, 72 p. AD 436 523.

46. O'Connor, J. Mechanized Indexing Methods and Their Testing. Journal of the Association of Computing Machinery, Vol. 11, No. 4, October 1964, pp. 437-449.

47. Oswald, V. A., Jr., et al. Automatic Indexing and Abstracting of the Contents of Documents, RADC-TR-59-208, prepared for Rome Air Development Center, ARDC, USAF, Oct. 1959, p. 133.

48. Ramo-Wooldridge, Division of Thompson Ramo Wooldridge, Inc. Final Report on the Study of Automatic Abstracting. September 1961. AD 269 600.

49. Ramo-Wooldridge Laboratories, Word Correlation and Automatic Indexing, Canoga Park, Calif. 1959.

50. Salton, G. A Combined Program of Statistical and Linguistic Procedures for Automatic Information Classification and Selection. Automation and Scientific Communication: American Documentation Institute, 26th Annual Meeting, Chicago, Illinois, October 7-11, 1963. American Documentation Institute, Washington, D. C., 1963, pp. 53-54.

51. Shannon, R. L. Experiment in Semiautomatic Indexing. NSAEC: Appended to Research and Development Abstracts of the NSAEC, July-September 1962.

52. Slamecka, V. and Zunde, P. Automatic Subject Indexing from Textual Condensations, Automation and Scientific Communication. Automation and Scientific Communication: ADI 26th Annual Meeting, Chicago, Illinois, October 7-11, 1963. American Documentation Institute, Washington, D. C., 1963.

53. Stevens, M. E. Preliminary Results of a Small-Scale Experiment in Automatic Indexing. NATO Advanced Study Institute on Automatic Document Analysis, Venice, Italy, 1963.

54. Stevens, M. E., Automatic Indexing: A State-of-the-Art Report, National Bureau of Standards Monograph 91, Washington, D. C., 1965.

55. Stevens, M. E. and Urban, G. H. Training a Computer to Assign Descriptors to Documents: Experiments in Automatic Indexing. AFIPS Conference Proceedings: Spring Joint Computer Conference, Washington, D. C., 1964, Vol. 25. Spartan Books, Baltimore, Md., 1964.

56. Swanson, D. R. Searching Natural Language Text by Computer. Science, Vol. 132, No. 3434, October 21, 1960, pp. 1099-1104.

57. Swanson, D. R. Research Procedures for Automatic Indexing. Institute on Information Storage and Retrieval. Machine Indexing: Progress and Problems, Papers Presented at the 3rd Institute, February 13-17, 1961. The American University, Washington, D. C., 1962, pp. 281-304.

58. Swanson, D. R. Automatic Indexing and Classification. NATO Advanced Study Institute on Automatic Document Analysis, Venice, Italy, 1963.

59. Trachtenberg, A. Automatic Document Classification Using Information Theoretical Methods. Automation and Scientific Communication, ADI 26th Annual Meeting, Chicago, Illinois, American Documentation Institute, Washington, D. C., 1963, pp. 349-350.

60. Turner, L. D. and Kennedy, J. H. System of Automatic Processing and Indexing of Reports. University of California, Lawrence Radiation Laboratory, Livermore, California, July 1961, p. 32.

61. Williams, J. H., Jr. A Discriminant Method for Automatically Classifying Documents. AFIPS Conference Proceedings: Fall Joint Computer Conference, 1963, Vol. 24. Spartan Books, Baltimore, Md., 1963, pp. 161-166.

B. SELECTED REFERENCES FROM RELATED FIELDS

62. Advance Information Systems Division, Hughes Dynamics, Inc., Sherman Oaks, Calif. Final Report on the Organization of Large Files. Part IV, April 1964.

63. Advanced Information Systems Company. Report on the Organization of Large Files with Self-Organizing Capability, p. 102

64. Agrayev, V. A. and Borodin, V. V. The Problem of Automatic Abstracting and Possible Solutions. Tezisy Soveshchaniya po Matematicheskoy Lingvistike (Abstracts of the Conference on Mathematical Linguistics, Leningrad, April 5-21, 1959, p. 49) 271 JPRS-893, p. 90.

65. Arthur D. Little, Inc., Cambridge, Mass. Automatic Message Retrieval, November 1963.

66. Automated File Storage and Retrieval Office. Automated File Storage and Retrieval, Vol. 53, May 1961, p. 140. AD 403 826.

67. Baker, F. B. Information Retrieval Based upon Latent Class Analysis. Journal of the Association for Computing Machinery, Vol. 9, No. 4, October 1962, pp. 512-521.

68. Bar-Hillel, Y. An Examination of Information Theory. Philosophy of Science, Vol. 22, April 1955, pp. 86-105.

69. Barnes, R. Language Problems Posed by Heavily Structured Data. Journal of the Association for Computing Machinery, Vol. 5, No. 1, January 1962.

70. Belevitch, V. On the Statistical Laws of Linguistic Distributions. Enseignement Preparatoire aux Techniques de la Documentation Automatique, EURATOM, Brussels, February 15-22, 1960.

71. Bell, D. A. and Ross, A. S. C. Negative Entropy of Welsh Words. Information Theory: Third London Symposium, September 12-16, 1955. Academic Press, New York, N. Y., 1956, pp. 149-153.

72. Bernshtein, E. S. Kvoprosy ob avtomaticheskom indeksirovanii (On the Problem of Automatic Indexing), Nauchno-Tekhnicheskaya Informazya (Scientific-Technical Information), No. 10, 1963, Moscow, pp. 22-25.

73. Bernstein, H. H. and Petri, H. Ein Mechanisches Verfahren Zur Herstellung Selektiver Listen Mit Hilfe Der Lochstreifentechnik. Das Rationelle Buro, 15 Jahrgang - Heft 3, 1964.

74. Bershadskiy, R. Y. Soviet Developments in Information Storage and Retrieval, December 1962. AD 400 597.

75. Black, D. V. Indexing Techniques: Description and Background. Document Storage and Retrieval Techniques. Planning Research Corporation, Los Angeles, Calif., June 1963, 29 p. AD 414 713, AD 417 452.

76. Blyth, C. R. Note on Estimation Information. Applied Mathematics and Statistics Laboratory, Stanford University, Stanford, Calif., January 1958, 18 p.

77. Bohnert, L. M. New Role of Machines in Document Retrieval: Definitions and Scope. Institute of Information and Storage Retrieval. Machine Indexing: Progress and Problems. Papers Presented at the 3rd Institute, Washington, D. C., February 13-17, 1961. The American University, Washington, D. C., 1961, pp. 8-11.

78. Borko, H. Evaluating the Effectiveness of Information Retrieval Systems. System Development Corporation, Santa Monica, Calif. September 1962. AD 288 835.

79. Bose, R. C. On Some Connections Between the Design of Experiments and Information Theory. Case Institute of Technology, University of North Carolina, 1961.

80. Bourne, C. P. Bibliography on the Mechanization of Information Retrieval. Stanford Research Institute, Menlo Park, Calif., February 1958. AD 10 151.

81. Bourne, C. P. Bibliography on the Mechanization of Information Retrieval: Supplement 1. Stanford Research Institute, Menlo Park, Calif., February 1959.

82. Bourne, C. P. Bibliography on the Mechanization of Information Retrieval: Supplement II. Stanford Research Institute, Menlo Park, Calif., February 1960, 14 p.

83. Bourne, C. P. Methods of Information Handling. John Wiley & Sons, Inc., New York, N. Y., 1963, 241 p.

84. Bourne, C. P. and Ford, D. F. A Study of the Statistics of Letters in English Words. Information and Control, Vol. 4, No. 1, March 1961, pp. 48-67.

85. Bushnell, D. and Borko, H. Information Retrieval Systems and Education. System Development Corporation, Santa Monica, Calif., Presented at APA Convention, St. Louis, Mo., September 1962.

86. Carlson, G. Search Strategy by Reference Libraries. Final Report on the Organization of Large Files. Part III. Advance Information Systems Division, Hughes Dynamics, Inc., Sherman Oaks, Calif., March 1964.

87. Carnap, R. and Bar-Hillel, Y. An Outline of a Theory of Sematic Information. Research Laboratory of Electronics, Massachusetts Institute of Technology, October 1952, 48 p.

88. Carter, L. F. and Marzocco, F. N. Technical Memorandum, January 1964. N64-16257.

89. Chavchanidze, V. V. and Kumsishvilli, V. A. The Determination of Laws of Distribution on the Basis of a Small Number of Observations. Primeneniye Vychislitel'noy Tekhniki dlya Avtomatizatsii Proizvodstva, Mashgiz, Moskva, 1961, pp. 129-139. AD 299 643.

90. Claridge, P. R. P., Mechanized Indexing of Information on Chemical Compounds in Plants, The Indexer, Vol. 2, No. 1, July 1961, pp. 178-183.

91. Collison, R. L. Indexes and Indexing. John Degraff, Inc., New York, N. Y., 1959, 200 p.

92. Dale, A. G. et al, A Programming System for Automatic Classification with Applications in Linguistic and Information Retrieval Research, Linguistics Research Center, University of Texas, Austin, Texas, October 1964, 19 p.

93. Datatrol Corporation, Silver Spring, Md. Datatrol 1401 ALP: A Generalized Information Retrieval Program for the IBM 1401 Computer. February 1963.

94. Defense Documentation Center, Alexandria, Va. Information Theory: Bibliography. Technical Abstract Bulletin, Vol. 16, August 15, 1964, p. IV. AD 269 800 (U).

95. DeLucia, A. Index-Abstract and Design. American Documentation, Vol. 15, No. 2, 1964, pp. 121-125.

96. Detant, M., Lecerf, Y., and Leroy, A. Travaux Pratiques sur Lietablissement des Diagrammes. Ensignement Preparatoire aux Techniques de la Documentation Automatique. EURATOM, Brussels, February 1960.

97. Detant, M. and Leroy, A. Elaboration d'un Programme D'Analyse de la Signification. EURATOM, Brussels, June 1961.

98. Directorate of Information Sciences, AFOSR, Information Sciences 1963, Annual Report, Washington, D. C., January 1964.

99. Douglas Aircraft Company. Missile and Space Systems Division, Santa Monica, Calif. Mechanized Information Retrieval System for Douglas Aircraft Company, Inc., January 1962.

100. Dowell, N. G. and Marshall, J. W. Experience with Computer-Produced Indexes. ASLIB Proceedings, Vol. 14, October 1962, pp. 323-332.

101. Doyle, L. B. Semantic Road Maps for Literature Searchers. System Development Corporation, Santa Monica, Calif., January 1961, 29 p. AD 284 259.

102. Doyle, L. B. Semantic Road Maps for Literature Searchers. Journal of American Computing Machinery, Vol. 8, No. 4, October 1961, pp. 553-578.

103. Doyle, L. B. Discussion of a Proposed Study of Associations Derived from Text. System Development Corporation, Santa Monica, Calif., December 1961, 11 p. AD 282 693.

104. Doyle, L. B. Indexing and Abstracting by Association. System Development Corporation, Santa Monica, Calif., 1962.

105. Doyle, L. B. Statistical Semantics. System Development Corporation, Santa Monica, Calif., 1962. AD 281 909.

106. Doyle, L. B. Indexing and Abstracting by Association. American Documentation, Vol. 13, No. 4, October 1962, pp. 378-390.

107. Doyle, L. B. The Microstatistics of Text. Information Storage and Retrieval, Vol. 1, No. 4, 1963, pp. 189-214.

108. Doyle, L. B. The Microstatistics of Text. System Development Corporation, Santa Monica, Calif., February 1963, 36 p.

109. Doyle, L. B. Some Compromise between Word Grouping and Document Grouping. System Development Corporation, Santa Monica, Calif., March 1964, 22 p. AD 440 044.

110. Engineers Joint Council, New York, N. Y. Study of Engineering Terminology and Relationships Among Engineering Terms. Final Report, August 1963.

111. Estrin, G. Maze Structure and Information Retrieval. Proceedings of the International Conference on Scientific Information, Washington, D. C., Vol. 2, 1959, pp. 1383-1393. AD 250 162.

112. Fairthorne, R. A. Mathematics, Mechanics, and Statistics for Information Science Curriculum or, What Mathematics Does and Information Scientist Need? Automation and Scientific Communication: ADI 26th Annual Meeting, Chicago, Illinois, October 7-11, 1963. American Documentation Institute, Washington, D. C., 1963.

113. Fano, R. M. Information Theory and Retrieval of Recorded Material. Documentation in Action. Shera, J. H., et al., Reinhold, N. Y., 1956, pp. 238-244.

114. Farrandane, J. Relational Indexing and Classification in Light of Recent Experimental Work in Psychology. Information Storage and Retrieval, Vol. 1, No. 1, March 1963, pp. 3-11.

115. Fleischer, H. An Introduction to the Theory of Information. Library Quarterly, Vol. 25, October 1955, pp. 326-332.

116. Flood, M. The Systems Approach to Library Planning. Library Quarterly, Vol. 34, No. 4, October 1964, pp. 326-338.

117. Frome, J. Semiautomatic Indexing and Encoding. Franklin Institute Journal, Vol. 270, No. 1, July 1960, pp. 3-26.

118. Frumkina, R. M. Methods of Comparing Statistical Dictionaries. Tezisy Soveshchaniya po Matematicheskoy Lingvistike (Abstracts of the Conference on Mathematical Linguistics, Leningrad, April 15-21, 1959, p. 7.) 229 JPRS-893, p. 20.

119. Fucks, W. Mathematical Theory of Word Formation. Information Theory: Third London Symposium, September 12-16, 1955. Academic Press, New York, N. Y., 1956, pp. 154-170.

120. Gilbert, P. T., Jr. An Optimal Punch Card Code for General Files. American Documentation, Vol. 9, No. 2, April 1958.

121. Giuliano, V. E. Automatic Message Retrieval by Associative Techniques. First Congress on the Information System Sciences, Hot Springs, Va., November 1962.

122. Giuliano, V. E. and Jones, P. E. Linear Associative Information Retrieval. Vistas in Information Handling. Spartan Books, Washington, D. C., 1963, pp. 30-54.

123. Goffman, W. Probalistic Models in Information. Information Retrieval in Action. Western Reserve University Press, Cleveland, Ohio, 1963, pp. 155-160.

124. Goldman, A. J., Bender, B. K., et al. Mathematical Research Related to Information Selection Systems. Final Report, National Bureau of Standards, Washington, D. C., June 30, 1961, 58 p. AD 262 882.

125. Gottesman, J. and Gottesman, E. Machines, Documentation, and Automation Coding. American Documentation, Vol. 8, No. 2, April 1957, pp. 129-133.

126. Grignetti, M. C. A Note on the Entropy of Words in Printed English. Information and Control, Vol. 7, No. 3, September 1964.

127. Gull, C. D. Guidelines to Mechanizing Information Systems. Information Retrieval Management. American Data Processing, Inc., 1962, pp. 101-110.

128. Hardwick, N. H., et al. Fact Correlation Experimentation, May 1964. AD 603 697.

129. Harlow, J. Research in Information Retrieval, July-September 1962. AD 400 569.

130. Harlow, J. Research in Information Retrieval, October-December 1962. AD 401 914.

131. Harlow, J. An Investigation of the Techniques and Concepts of Information Retrieval, Report No. 5. ITT Federal Electric Corporation, Paramus, N. J., September 1963, 47 p. AD 429 837.

132. Harlow, J. An Investigation of the Techniques and Concepts of Information Retrieval, Report No. 6. ITT Data and Information Systems Division, Paramus, N. J., January 31, 1964, 91 p. AD 437 924.

133. Harlow, J., Trachtenberg, A., Darmstadt, D., Greenberg, G., and Szejman, A. Research in Information Retrieval. ITT Data and Information Systems Division, Paramus, N. J., January 1963.

134. Harlow, J., Trachtenberg, A., Darmstadt, D., Greenberg, G., and Szejman, A. Research Information Retrieval. ITT Data and Information Systems Division, Paramus, N. J., April 1963.

135. Harman, H. H. Modern Factor Analysis. University of Chicago Press, Chicago, Illinois, 1960.

136. Harris, B. Determining Bounds on Integrals with Applications to Cataloguing Problems. Stanford University Applied Mathematics and Statistics Laboratory, Stanford University, Palo Alto, Calif., April 1958, 56 p. AD 158 558.

137. Hattery, L. H. and McCormick, E. M., Editors. Information Retrieval Management. American Data Processing, Inc., Detroit, Mich., 1962, 200 p.

138. Hays, D. G. Report of a Summer Seminar on Computational Linguistics, February 1964. DDC 431 868.

139. Heilprin, L. B. Mathematical Model of Indexing. Documentation Incorporated, Bethesda, Md., August 1957. AD 136 477.

140. Heilprin, L. B. Toward a Definition of Information Science. Automation and Scientific Documentation; ADI 26th Annual Meeting, Chicago, Illinois, October 7-11, 1963. American Documentation Institute, Washington, D. C., 1963.

141. Hensley, C. B. Selective Dissemination of Information: State-of-the-Art in May 1963. AFIPS Conference Proceedings; Spring Joint Computer Conference, Detroit, Mich., May 1963, Vol. 23. Spartan Books, Baltimore, Md., 1963, pp. 257-262.

142. Herner, S. Deciding When to Establish Your Own Storage and Retrieval System. Herner and Company, Washington, D. C. AD 432 518.

143. Hetrick, J. M. Distribution Function in Documentation. Special Libraries, Vol. 50, May 1959, pp. 193-195.

144. Hillman, D. J. Study of Theories and Models of Information Storage and Retrieval. Report No. 3: A Positive Model for Systems of Special Classification. Center of Information Sciences, Lehigh University, Bethlehem, Pa., August 1962, 20 p. AD 283 640.

145. Hillman, D. J. Study of Theories and Models of Information Storage and Retrieval: New Foundations for Retrieval Theories. Report No. 4. Center for Information Sciences, Lehigh University, Bethlehem, Pa., August 12, 1963, 19 p.

146. Hillman, D. J. Study of Theories and Models of Information Storage and Retrieval: Positive Models of Retrieval Systems as Species of Logical Algebras. Report No. 5. Center for the Information Sciences, Lehigh University, Bethlehem, Pa., 1963, 24 p.

147. Hillman, D. J. Study of Theories and Models of Information Storage and Retrieval: Retrieval Systems for Non-Static Document Collections. Report No. 6. Center for the Information Science, Lehigh University, Bethlehem, Pa., September 26, 1963, 64 p.

148. Himmelman, D. S. and Chu, J. T. An Automatic Abstracting Program Employing Stylo-Statistical Techniques and Hierarchical Data Indexing. Association for Computing Machinery: 16th National Meeting. Reprint of Papers Presented. New York, N. Y., 1961.

149. Hockett, C. F. Review of the Mathematical Theory of Communication by Claude L. Shannon and Warren Weaver. Language, Vol. 29, No. 1, 1953, pp. 69-93.

150. Houston, N. and Wall, E. The Distribution of Term Usage in Manipulative Indexes. American Documentation, Vol. 15, No. 2, 1964, pp. 105-114.

151. Hirayama, K. Length of Abstract and Amount of Information. Journal of Chemical Documentation, Vol. 4, No. 1, 1963, pp. 9-11.

152. Hughes Dynamics Inc. The Organization of Large Files. Part I-VI, Advance Information System Division, Hughes Dynamics, Inc., Sherman Oaks, Calif., April 1964.

153. Institute for Cooperative Research, University of Pennsylvania, Mechanized Indexing: Some General Remarks and Some Small-Scale Empirical Results, Nov. 1960.

154. International Business Machines, White Plains, N. Y. Index Organization for Information Retrieval, 1961, 63 P.

155. Isert, I. L. Selective Retrieval From a Variable Length Information File, January 1962. AD 432 312.

156. Iung, J. Common Points between the Problems of Automatic Translation and Automatic Documentation. Ensignement Preparatoire aux Techniques de la Documentation Automatique, EURATOM, Brussels, February 15-22, 1960. 560 JPRS-8938.

157. Jackson, K. I. A Study of the Application of Present and Future Methods of Automation, Retrieval, and Portroyal to Department of Defense and NASA Engineering Documentation Systems and Centers. University of Alabama, University, Ala., 1963, 221 p. AD 417 583.

158. Jonker, F. Outline of a General Theory of Index Terminology and Indexing Methods. Jonker Business Machines, Inc., Gaithersberg, Md., October 1961, 47 p. AD 272 520.

159. King, D. W. Design of Experiments in Information Retrieval. Proceedings of the Social Statistics Section: American Statistical Association, Washington, D. C., 1963, pp. 103-118.

160. Kirsch, R. A. The Application of Automata Theory to Problems in Information Retrieval (with Selected Bibliography). National Bureau of Standards, Washington, D. C., March 1963, 70 p.

161. Klein, S. and Simmone, R. E. Automatic Analysis and Coding of English Grammar for Information Processing Systems. System Development Corporation, Santa Monica, Calif., 1962.

162. Kochen, M. High-Speed Document Perusal. International Business Machines Corporation, Yorktown Heights, N. Y., 1962. AD 285 255.

163. Kochen, M. Problems in Information Science with Emphasis on Adaptation to Use through Man-Machine Interaction. Final Report. I. J. Watson Research Corporation, Yorktown Heights, N. Y., 1963, Vol. I. AD 600 113.

164. Kogen, Kh. M. The Use of the Bull Punch Card Machines in Compiling Issue-by-Issue and Annual Indexes to Abstract Journals. Doklady Na Konferentsii Po Obrabotke Informatsii, Mashinnomu Perevodu I Avtomaticheskomu Chteniyu Teksta (Reports of the Conference on Information Processing, Machine Translation, and Automatic Text Reading). Institute of Scientific Information of the Academy of Sciences, USSR, Moscow, 1961, No. 7, pp. 1-14. 712 JPRS-13253.

165. Kohlstedt, D. W. and Markland, F. R. Grand Rapids Public Library Explores Mechanical Cataloguing. Library Journal, vol. 75, March 1950, pp. 417-418.

166. Koutsondas, A. M., Machol, R. E., and Minty, G. T. Frequency of Occurrence of Words: A Study of Zipf's Law, with Application to Mechanical Translation. University of Michigan, Willow Run Laboratories, University of Michigan, Ypsilanti, Mich., June 1957, 13 p.

167. Kraft, D. H. A Comparison of Keyword-in-Context (KWIC) Indexing of Titles with a Subject Heading Classification System. American Documentation, Vol. 15, No. 1, 1964, pp. 48-52.

168. Kreithen, A. Mathematical Foundations for a Storage and Retrieval Theory. Documentation Inc., Bethesda, Md., June 1957, 16 pp. AD 132 475.

169. Landauer, W. I. The Three as a Stratagem for Automatic Information Handling. More School of Electrical Engineering, University of Pennsylvania, Philadelphia, Pa., 1962, 121 pp. AD 293 888.

170. Lecert, Y. and Leroy, A. A Self-Priming Automatic Analysis Unit, Report No. 2. GRISA (Group for Research on Automatic Scientific Information), May 1960, p. 4. 576 JPRS-10 367.

171. Lefkovitz, D. and Prywes, N. S. Automatic Stratification of Information. AFIPS Conference Proceedings: Spring Joint Computer Conference, Detroit, Mich., 1963, Vol. 23. Spartan Books, Baltimore, Md., 1963, p. 413.

172. Leroy, A. Fully Automatic Documentation. Abstract from Enseignement Preparatoire aux Techniques de la Documentation Automatique, EURATOM, Brussels, February 15-22, 1960, p. 48. 561 JPRS-8938.

173. Livant, W. P. On the Occurrence of Preversible Words. American Documentation, Vol. 14, No. 3, 1963, pp. 234-237.

174. Luhn, H. P. Review of Information Retrieval Methods. IBM Corporation, Yorktown Heights, N. Y., 1958, 10 p.

175. Luhn, H. P. A Business Intelligence System. Journal of Research and Development, Vol. 2, No. 4, October 1958, pp. 314-319.

176. Luhn, H. P. Selective Dissemination of New Scientific Information with the Aid of Electronic Processing Equipment. IBM Corporation, Yorktown Heights, N. Y., 1959.

177. Luhn, H. P. Machinable Bibliographic Records as a Tool for Improving Communication of Scientific Information. Paper Prepared for Presentation at the 10th Pacific Scientific Congress, Honolulu, August 21 through September 6, 1961. IBM Corporation, White Plains, N. Y., 1961.

178. Luhn, H. P. Automated Intelligence System -- Some Basic Problems and Prerequisites for Their Solution. The Clarification, Unification, and Integration of Information Storage and Retrieval Proceedings (Symposium). New York City, N. Y., February 23, 1961.

179. Luhn, H. P. Automated Intelligence Systems. Information Retrieval Management. 1962 ed. Hattery, L. H. and McCormick, E. M., American Data Processing Inc., Detroit, Mich., 1962, pp. 92-100.

180. Luhn, H. P., Editor. Automation and Scientific Communication: 26th Annual Meeting, Chicago, Illinois, October 7-11, 1963. American Documentation Institute, Washington, D. C., 1963.

181. MacKay, D. M. The Place of "Meaning" in the Theory of Information. Information Theory: Third London Symposium, September 12-16, 1955, Academic Press, New York, N. Y., 1956, pp. 215-225.

182. Magnavox Research Laboratories, Torrence, Calif. Mathematical Models for Information Systems Design and a Calculus of Operations. Final Report, October 27, 1961, 178 p. AD 266 577.

183. Maizell, R. E. Value of Titles for Indexing Purposes. Revue de la Documentation, Vol. 27, August 1960, pp. 126-127.

184. Maloney, C. J. Semantic Information. American Documentation, Vol. 13, No. 3, 1962, pp. 276-288.

185. Mandelbrot, B. On the Theory of Word Frequencies and on Related Markovian Models of Discourse. Proceedings of Symposia on Applied Mathematics; Symposium on the Structure of Language and Its Mathematical Aspects. Vol. 12, March 1961, pp. 190-210.

186. Manly, R. Detailed Discussion of Bar-Hillel's "Theoretical Aspects of the Mechanization of Literature Searching." American Documentation, Vol. 15, No. 2, 1964, pp. 126-131.

187. Marill, T. M. Combinatorial Aspects of Information Retrieval. international Business Machines, Kingston, N. Y., November 1960, 9 p.

188. Maron, M. E. Mechanized Interpretation: the Logic Behind a Probabilistic Interpretation. Rand Corp., Santa Monica, Calif., April 1964, 20 p. AD 437 781.

189. Maron, M. E. and Kuhns, J. L. On Relevance, Probabilistic Indexi g, and Information Retrieval. Journal of the Association for Computing Machinery, Vol. 7, No. 3, July 1960, pp. 216-244.

190. Maron, M. E., Kuhns, J. L., and Ray, L. C. Probabilistic Indexing. A Statistical Technique for Document Identification and Retrieval. Thompson-Ramo-Wooldridge, Inc., Los Angeles, Calif., 1959, 91 p. AD 272 572.

191. Meetham, A. R., Probabilistic Pairs and Groups of Words in a Text, Language and Speech, Vol. 7, Pt. 2, April-June 1964, pp. 98-106.

192. Metcalf, J. W. Information Indexing and Subject Cataloging: Alphabetical, Classified, Coordinate, Mechanical. Scarecrow Press, New York, N. Y., 1957, 338 p.

193. Meyer-Uhlenreid, H. K. Automatisierung der Dokumentation und der Information in der Cetis der EURATOM. Nachrichten fur Dokumentation, Vol. 12, No. 1, March 1961, pp. 6-10.

194. Mitre Corporation, Bedford, Mass. First Congress on the Information System Sciences Session 4, Joint Man-Computer Decision Processes, February 1964, p. 111. AD 432 169.

195. Montgomery, C. and Swanson, D. R. Machinelike Indexing by People. American Documentation, Vol. 13, October 1962, pp. 359-366.

196. Mooers, C. N. Some Mathematical Fundamentals of the Use of Symbols in Information Retrieval. Zator Company, Cambridge, Mass., April 1959, 22 p. AD 213 782.

197. Needham, R. H. A Method for Using Computers in Information Classification. Information Processing (Proceedings of IFIP Congress, Munich). North-Holland Publishing Company, Amsterdam, 1962, pp. 284-287.

198. Newell, A. New Areas of Application of Computers. The Rand Corporation, November 1960. AD 432 326.

199. North American Aviation Space and Information Systems Division. Information Retrieval: Systems and Technology. A Literature Survey. 1951-1961. AD 403 826.

200. Nuyl, T. W. The "L'Unite" Mechanized Documentation. Revue de la Documentation, Vol. 28, No. 4, November 1961, pp. 140-147.

201. O'Connor, J. Correlation of Indexing Headings and Title Words in Three Medical Indexing Systems. American Documentation, Vol. 15, No. 2, April 1964, pp. 96-104.

202. Painter, A. F. An Analysis of Duplication and Consistency of Subject Indexing Involved in Report Handling at the Office of Technical Services, United States Department of Commerce. Rutgers State University, PhD. Dissertation. Office of Technical Services, Washington, D. C., 1963.

203. Perry, J. W. Characteristics of Recorded Information. Documentation in Action, by Shera, J. H., et al., Reinhold, N. Y. 1956, pp. 68-100.

204. Perry, J. W., Kent, A., and Berry, M. M. New Indexing-Abstracting System for Formal Reports, Development and Proof Services. Battelle Memorial Institute, Columbus, Ohio, 1955-1959. AD 125 030, AD 125 111, AD 125 112, AD 125 113.

205. Pevzner, B. R. and Styazhkin, N. I. A Method of Special Abstracting. Doklady Na Konferentsii Po Obrabotke Informatsii, Mashinnomu Perevodu I Avtomaticheskomu Chteniyu Teksta (Reports of the Conference on Information Processing, Machine Translation, and Automatic Text Reading). Institute of Scientific Information of the Academy of Sciences USSR, Moscow, 1961. 675 JPRS-13057.

206. Picot, G., Deribere-Desgardes, M. L. and Levery, F. Une Experience de Selection Automatique de Documentation. (An Experiment with the Automatic Selection of Documents). Revue Internationale de la Documentation, Vol. 29, 1962, pp. 8-13.

207. Pietsch, E. Mechanische Selektion: Summary of Paper at General Assembly of FID, September 22-25, 1959. Revue de la Documentation, Vol. 26, No. 66, August 1959.

208. Purto, V. A. Automatic Abstracting based on a Statistical Analysis of the Text. Doklady Na Konferentsii Po Obrabotke Informatsii, Mashinnomu Perevodu I Avtomaticheskomu Chteniyu Teksta (Reports of the Conference on Information Processing, Machine Translation, and Automatic Text Reading). Institute of Scientific Information of the Academy of Sciences USSR, Moscow, No. 9, 1961, pp. 1-16 706 JPRS-13196.

209. Quastler, H. A Primer on Information Theory. Controls Systems Laboratory, University of Illinois, Urbana, Illinois, January 1956, 66 p.

210. Raisig, L. M. Mathematical Evaluation of the Science Serial. Science, Vol. 131, No. 3411, May 1960, pp. 1417-1419.

211. Rath, G. J., Resnick, A., and Savage, T. R. The Formation of Abstracts by the Selection of Sentences. Part I. Sentence Selection by Men and Machines. American Documentation, Vol. 12, No. 2, 1961, pp. 139-141.

212. Resnick, A. The Formation of Abstracts by the Selection of Sentences. Part II. The Reliability of People in Selecting Sentences. American Documentation, Vol. 12, No. 2, April 1961, pp. 141-143.

213. Resnick, A. and Savage, T. R. The Consistency of Human Judgments of Relevance. American Documentation, Vol. 15, No. 2, 1964, pp. 93-95.

214. Ritchie, D. J. and Gottesman, N. H. The Mathematical Foundations of the Theory of Information. Bendix Aviation Corporation, Detroit, Mich., 1953, 163 p. AD 63 074.

215. Ruterbusch, M. J. Miracode--A New Break-Through in Automated Information Retrieval. Proceedings; Convention of the National Microfilm Association, San Francisco, Calif., April 30-May 2, 1963. Vol. 2 National Microfilm Association, Annapolis, Md., 1963, 367 p.

216. Saint-Gobain Company, Paris, France. General Information Manual and Information Retrieval Experiment. IBM Technical Publication Department, White Plains, N. Y., 1962, 28 p.

217. Salton, G. Information Storage and Retrieval. Computation Laboratory, Harvard University, Cambridge, Mass., 1961, 152 p. AD 274 816.

218. Salton, G. The Identification of Document Content: A Problem in Automatic Information Retrieval. Proceedings of a Harvard Symposium on Digital Computers and their Applications. April 1961. Computation Laboratory, Harvard University, Cambridge, Mass., 1962.

219. Salton, G. Some Experiments in the Generations of Word and Document Associations. AFIPS Conference Proceedings: Fall Joint Computer Conference, 1962, Vol. 22, pp. 234-250.

220. Salton, G. The Manipulation of Trees in Information Retrieval. Journal of the Association for Computing Machinery, Vol. 5, No.2, February 1962.

221. Salton, G. Information Storage and Retrieval. Computation Laboratory, Harvard University, Cambridge, Mass., 1963, 144 p. AD 408 934.

222. Salton, G. Some Hierarchical Models for Automatic Document Retrieval. American Documentation, Vol. 14, No. 3, July 1963, pp. 213-222.

223. Salton, G. Associative Document Retrieval Techniques Using Bibliographic Information. Journal of the Association for Computing Machinery, Vol. 10, No. 4, October 1963, pp. 440-457.

224. Salton, G., et al. Information Storage and Retrieval. Computation Laboratory, Harvard University, Cambridge, Mass., 1962. AD 287 945.

225. Salton, G. and Sussenguth, E. H., Jr. Automatic Structure Matching Procedures and Some Typical Retrieval Applications. Computation Laboratory, Harvard University, to Air Force Cambridge Research Laboratories, Cambridge Mass.

226. Samson, E. W. Theory of Information: The Basic Theorems on System Uncertainty. Air Force Cambridge Research Center, Cambridge, Mass., November 1953, 25 p.

227. Samson, E. W. Information Theory: Questions and Uncertainties. Air Force Cambridge Research Center, Cambridge, Mass., January 1954, 45 p. AD 28 403.

228. Schouten, J. F. Ignorance, Knowledge and Information. Information Theory: Third London Symposium, September 12-16, 1955. Academic Press, New York, N. Y., 1956, pp. 37-46.

229. Schutzenberg, M. P. On Some Measures of Information Used in Statistics. Information Theory: Third London Symposium, September 12-16, 1955. Academic Press, New York, N. Y., 1956, pp. 18-25.

230. Schutzenberger, M. Information Theory. Abstract of Papers from Enseignement Preparatoire Aux Techniques de la Documentation Automatique (Preliminary Study on the Techniques of Automatic Documentation). European Atomic Energy Community, EURATOM, Brussels, February 15-22, 1960.

231. Schwartz, E. S. A Dictionary for Minimum Redundancy Encoding. Journal of the Association for Computing Machinery, Vol. 10, No. 4, 1963, pp. 413-439.

232. Sebeok, T. A. Information Structures in Linguistics, December 1963. AD 433 001.

233. Shannon, C. E. The Mathematical Theory of Communication. Bell System Technical Journal, Vol. 27, 1948, pp. 379-423, 623-658.

234. Shaw, R. R. Mechanical Storage, Handling, Retrieval, and Supply of Information. Advisory Group for Aeronautical Research and Development, Paris, France, February 1956, 34 p. AD 144 216.

235. Shera, J. H., et al. Areas for Research. Documentation in Action, Reinhold, N. Y., 1956, pp. 447-458.

236. Shera, J. H., Kent, A., and Perry, J. W. Information Systems in Documentation. Interscience Publishers, New York, N. Y., 1957, 639 p.

237. Shoffner, R. M. Self-Organizing Processes--Computer Programs for Files of Bit Patterns. Final Report on the Organization of Large Files, Part IV. Advance Information Systems Division, Hughes Dynamics, Inc., Sherman Oaks, Calif., April 1964.

238. Sievers, P. T. and Fasana, P. J. Automated Routines in Technical Services. Presented at the 7th Military Libraries' Workshop October 2-4, 1963, Silver Spring, Md. DDC 435 615.

239. Simmons, R. F. and McConlogue, K. L. Maximum-Depth Indexing for Computer Retrieval of English Language Data. System Development Corporation, Santa Monica, Calif., April 1962, 22 p. AD 275 814.

240. Simmons, R. F., Klein, S., and McConlogue, K. Co-occurrence and Dependency Logic for Answering English Questions. System Development Corporation, Santa Monica, Calif., April 1963.

241. Sinnett, J. D. An Evaluation of Links and Roles Used in Information Retrieval, December 1963. AD 432 198.

242. Smullyan, R. M. Theory of Formal Systems, Princeton University Press, Princeton, N. J., 1961, 142 p.

243. Spiegel, J., Bennett, E., Haines, E., Vicksell, R., and Baker, J. Statistical Association Procedures for Message Content Analysis. Mitre Corporation, Bedford, Mass., October 1962.

244. Stearns, S. D. Using Bayes' Induction Theorem to Estimate Probabilities in a Communication Channel. The Dikewood Corporation, Albuquerque, New Mexico, July 1961. AD 265 847.

245. Stiles, H. E. The Association Factor in Information Retrieval. Journal of the Association for Computing Machinery, Vol. 8, No. 2, April 1961, pp. 271-279.

246. Stouffer, S. A., Lazarsfeld, P. F., et al., Measurement and Prediction, Princeton University Press, 1950.

247. Styazhkin, N. I. Basic Trends in Modern Documentation and Possibilities of Constructing Mathematical Logic Theories of Information Search Systems. Soobshcheniya Laboratorii Elektro modelirovaniya (Information on Laboratory Electrical Model Operation). Institute of Scientific Information of the Academy of Sciences USSR, Moscow, No. 1, 1960, pp. 1-250.

248. Swanson, D. R. Interrogating the Computer in Natural Language. International Federation of Information Processing Societies Second Congress, Munich, 1962, p. 124-127.

249. Swets, J. A. Information Retrieval Systems. Science, Vol. 141, July 19, 1963, pp. 245-250.

250. Tanimoto, T. T. An Elementary Mathematical Theory of Classification and Prediction. International Business Machines Corporation, New York, N. Y., 1958.

251. Tanimoto, T. T. The General Problem of Classification and Indexing. Machine Indexing: Progress and Problems. Papers Presented at the Third Institute on Information Storage and Retrieval, February 13-17, 1961. The American University, Washington, D. C., 1962, pp. 233-235.

252. Taylor, A., et al. Quantitative Methods for Information Processing Systems Evaluation, January 1964. Auerback Corporation, Philadelphia, Pa., AD 435 557.

253. Thompson Ramo Wooldridge, Inc. The Study for Automatic Abstracting. Canoga Park, Calif., September 1961. AD 269 599, AD 269 600.

254. Thurstone, L. L., Multiple-Factor Analysis, University of Chicago Press, Chicago, Ill., 1947.

255. Toll, M. G. System Components Information Center, September 1963. DDC 425 581.

256. Touloukian, Y. S. The Concept of Entropy in Communication, Living Organisms, and Thermodynamics. Purdue University, Engineering Experiment Station, Lafayette, Indiana, 1956, 66 p.

257. Trachtenberg, A., et al. Investigation of the Techniques and Concepts of Information Retrieval. Progress Report. ITT Federal Electric Corporation, Paramus, N. J., January-March 1963. AD 413 620.

258. Trachtenberg, A., Darmstadt, Q. A., and Greenberg, G. An Investigation of the Techniques and Concepts of Information Retrieval. Progress Report No. 2, October-December 1962. ITT Federal Electric Corporation, Paramus, N. J., 1963, 78 p. AD 401 914.

259. TRW Computers Company. Automatic Abstracting: Final Report. Canoga Park, Calif., February 1963, 48 p. AD 406 155.

260. U. S. Congress Senate Committee on Government Operations, 87th Congress, First Session. Documentation, Indexing, and Retrieval of Scientific Information, Supplement. Washington, D. C., 22 p. Senate Document No. 15.

261. U. S. Congress Senate Committee on Government Operations, 86th Congress, Second Session. Documentation, Indexing, and Retrieval of Scientific Information. Washington, D. C., 1961, 283 p. Senate Document No. 113.

262. U. S. Department of Commerce - Office of Technical Services. Information Storage and Retrieval. Washington, D. C., September 1961, 13 p.

263. U. S. National Library of Medicine. National Library of Medicine Index Mechanization Project, July 1, 1958-June 30, 1960. Medical Library Association Bulletin, Vol. 49, No. 1 (Part II). January 1961, pp. 1-96.

264. Uskavick, C. W. Concepts of Automatic Data Storage and Retrieval in the Simplex System. Massachusetts Institute of Technology, Lincoln Laboratories, Lexington, Mass., October 1960, 67 pp. AD 245 472, PB 152 787.

265. Verhoeff, J., Goffman, W., Belzer, J. On the Inefficiency of Boolean Functions in Information Retrieval. Communications of the ACM, Vol. 4, December 1961, pp. 10-12.

266. Vickery, B. C. Developments in Subject Indexing. Journal of Documentation, Vol. 11, March 1955, pp. 1-11.

267. Vickery, B. C. Problems in the Construction of Information Retrieval Systems. Journal of Documentation, Vol. 14, No. 3, September 1958, pp. 136-143.

268. Vickery, B. C. The Statistical Method in Indexing. Revue de la Documentation, Vol. 28, No. 2, 1961, pp. 56-62.

269. Wadsworth, H. M. and Booth, R. E. Some Statistical Sampling Concepts Applied to the Information Retrieval Process of Documentation Systems. Western Reserve University, Center for Documentation and Communication Research, Cleveland, Ohio, August 1958, 38 p. AD 201 864.

270. Wadsworth, H. M. and Booth, R. E. The Application of Statistical Decision Theory to Problems of Documentation. Western Reserve University, Center for Documentation and Communication Research, Cleveland, Ohio, March 1959, 24 p.

271. Walkowicz, J. L. A Bibliography of Foreign Developments in Machine Translation and Information Processing. National Bureau of Standards, Washington, D. C., July 10, 1963, 191 p.

272. Warheit, I. A. The Direct Access Search System. AFIPS Conference Proceedings: Fall Joint Computer Conference, 1963, Vol. 24. Spartan Books, Baltimore, Md., 1963, pp. 167-172.

273. Warheit, I. A. The Use of Computers in Information Retrieval. Information Retrieval Today, 1963. Minneapolis, Minnesota Center for Continuation Study, University of Minnesota, 1963, pp. 55-69.

274. Watanabe, S. A Note on the Formation of Concept and of Association by Information-Theoretical Correlation Analysis. Information and Control, Vol. 4, 1961, pp. 291-296.

275. Western Reserve University. Center for Documentation and Communication Research. Documentation and Information Retrieval. A Selected Bibliography. Cleveland, Ohio, 1961, 8 p.

276. Western Reserve University Press, Cleveland, Ohio. Information Retrieval In Action. (A collection of 25 papers presented to a conference). 1963, 310 p.

277. Wiegand, K. L. Information Theory and Human Behavior: Uncertainty as a Fundamental Variable in Information-Processing Tasks, October 1963. DDC 423 557.

278. Wright, M. A. Matching Inquiries to an Index. The Computer Journal, Vol. 4, No. 1, April 1961, pp. 38-41.

279. Wyllys, R. E. Research in Techniques for Improving Automatic Abstracting Procedures. System Development Corporation, Santa Monica, Calif., April 1963, 30 p. AD 404 105.

280. Yngve, V. H. COMIT as an IR Language. Communications of the ACM, January 1962, pp. 19-28.

281. Zunde, P., Some Mathematical Aspects of Automatic Indexing, MS Thesis, George Washington University, June 1965, 113 p.

Unclassified
Security Classification

DOCUMENT CONTROL DATA - R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)	2a. REPORT SECURITY CLASSIFICATION
Documentation, Inc., Bethesda, Maryland	Unclassified
	2b GROUP

3. REPORT TITLE

Automatic Indexing from Machine Readable Abstracts of Scientific Documents (FAST).

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

Scientific Report

5. AUTHOR(S) (Last name, first name, initial)

Zunde, Pranas

6. REPORT DATE	7a TOTAL NO OF PAGES	7b. NO. OF REFS
September, 1965	210	282

8a. CONTRACT OR GRANT NO.	9a ORIGINATOR'S REPORT NUMBER(S)
AF-49(604)-4236	
b. PROJECT NO. 110	AFOSR 65-1425
c.	9b OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
d.	

10. AVAILABILITY/LIMITATION NOTICES

N/A Each transmittal of this document outside the agencies of the U. S. Government must have prior approval of AFOSR/SRGL.

11. SUPPLEMENTARY NOTES	12 SPONSORING MILITARY ACTIVITY
	Air Force Office of Scientific Research

13 ABSTRACT

State-of-the-art of machine indexing is reported. Various proposed machine indexing methods are reviewed and evaluated. Methods for comparing machine and human indexing as well as machine indexing systems among themselves are described. Possible approaches to various problem solutions in machine indexing are indicated.

Part Two of the report describes the design of the Formal Autoindexing of Scientific Texts (FAST) system. Characteristics of Uniterm co-ordinate indexes are investigated and generalizations to scientific indexes made. Laws for the formation of words in the indexing language are derived and verified. The operational principles of the FAST system and test results of various system components are reported. Indexes produced by the FAST method are compared with those produced by human indexers for inter-indexer and intra-indexer consistency. A method of formal evaluation of indexes using the information theory approach is presented and applied to the FAST and conventional indexes. It is concluded that the FAST system can produce Uniterm co-ordinate indexes adequate to user's requirements better and faster than human indexers can do.

DD FORM 1 JAN 64 1473

Unclassified
Security Classification

Unclassified
Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Abstract Autoindexing Consistency Evaluation FAST Indexing Programming State-of-the-art Testing Uniterm						

INSTRUCTIONS

1. **ORIGINATING ACTIVITY:** Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization *(corporate author)* issuing the report.

2a. **REPORT SECURITY CLASSIFICATION:** Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2b. **GROUP:** Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. **REPORT TITLE:** Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. **DESCRIPTIVE NOTES:** If appropriate, enter the type of *report, e.g., interim, progress, summary, annual, or final.* Give the inclusive dates when a specific reporting period is covered.

5. **AUTHOR(S):** Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. **REPORT DATE:** Enter the date of the report as day, month, year, or month, year. If more than one date appears on the report, use date of publication.

7a. **TOTAL NUMBER OF PAGES:** The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7b. **NUMBER OF REFERENCES:** Enter the total number of *references cited in the report.*

8a. **CONTRACT OR GRANT NUMBER:** If appropriate, enter the applicable number of the contract or grant under which the report was written.

8b, 8c, & 8d. **PROJECT NUMBER:** Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9a. **ORIGINATOR'S REPORT NUMBER(S):** Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9b. **OTHER REPORT NUMBER(S):** If the report has been assigned any other report numbers *(either by the originator or by the sponsor),* also enter this number(s).

10. **AVAILABILITY/LIMITATION NOTICES:** Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through _____."

(4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through _____."

(5) "All distribution of this report is controlled. Qualified DDC users shall request through _____."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. **SUPPLEMENTARY NOTES:** Use for additional explanatory notes.

12. **SPONSORING MILITARY ACTIVITY:** Enter the name of the departmental project office or laboratory sponsoring *(paying for)* the research and development. Include address.

13 **ABSTRACT:** Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as *(TS), (S), (C), or (U).*

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14 **KEY WORDS:** Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, rules, and weights is optional.

Unclassified
Security Classification